DOCUMENT RESUME

ED 040 051 SE 008 294

AUTHOR	Callahan, John; And Others
TITLE	Modern Mathematics for Elementary Teachers: A Laboratory Approach
INSTITUTION	Cambridge Conference on School Mathematics, Newton, Mass.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	69
NOTE	280p.
EDRS PRICE	EDRS Price MF-$1.25 HC Not Available from EDRS.
DESCRIPTORS	*Curriculum Development, *Elementary School Teachers, Geometric Concepts, *Instructional Materials, Laboratory Manuals, Mathematics, *Modern Mathematics, *Teacher Education
IDENTIFIERS	Cambridge Conference on School Mathematics

ABSTRACT

Reports on the development of a sample course in modern mathematics for elementary teachers. The approach to the course was based on three methodological principles--(1) that emphasis should be placed on mathematics as an organization of (experimental and other) information and not primarily as a deductive system, (2) that it is important to use concrete objects to emphasize the "real " nature of mathematics, and (3) that the material taught in the course should have a direct bearing on the material that the prospective teacher would use in the elementary school classroom. The subject matter of the course consists of a study of the positive real numbers in connection with the measurement process, the study of the whole number line as a one dimensional vector space, and the study of vector geometry in the plane. [Not available in hardcopy due to marginal legibility of original document.] (RP)

ED040051

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

MODERN MATHEMATICS FOR ELEMENTARY TEACHERS

A LABORATORY APPROACH

by

Professor John Callahan
Department of Mathematics
Boston State College

Professor Shlomo Sternberg
Department of Mathematics
Harvard University

Professor Edwin Weiss
Department of Mathematics
Boston University

1969

Cambridge Conference on School Mathematics

Funded by

Undergraduate Science Curriculum Improvement Division

National Science Foundation

SE 008 294

Callahan/Sternberg/Weiss
April 1, 1969

Modern Mathematics for Elementary Teachers

A Laboratory Approach

Introduction i-iii

Chapter I Measurement

Laboratory Manual for Chapter I

Shlomo Sternberg
December 1967

Boston State College

In accordance with the recommendations of the Cambridge Conference meeting of the summer of 1966, Professor Weiss and I embarked on the project of developing a sample course for use in the teacher's college. We worked in close cooperation with Professors Perrault and Callahan of the Boston State College. From the beginning we felt strongly that it was important to develop the material in close cooperation with a teacher's college. The reason for this was that our first main problem was one of educating ourselves as to the nature of the students and educational atmosphere in the teacher's college. During the fall semester our operating procedure was as follows: We would hold weekly (or bi-weekly) meetings with Professor Callahan (who was teaching the course) to discuss educational objectives and methods prior to classes. I then attended the class (disguised as a student registered in the class) and Professor Weiss visited the class at regular intervals. (Actually, my schedule allowed me to attend only two of the three classes per week and Professor Weiss visited the third class.) By being "part of the class" I was able to get to know the reactions of a few of the students quite well. In retrospect I can say that this procedure was extremely valuable. I learned a great deal about the nature of the educational problems and this information was used in revising the material for the course.

Our approach to the course was based on three methodological principles: 1) that emphasis should be placed on mathematics as an organization of (experimental and other) information and not primarily as a deductive system; 2) that it is important to use concrete objects to emphasize the "real" nature of mathematics; 3) that the material taught in the course should have a direct bearing on the material that the prospective teacher would use in the elementary school classroom.

As to 1): There is no doubt that the key feature distinguishing mathematics from the other sciences is its purely deductive character. However, it is our feeling that (especially with the students under question) this point has been over-emphasized at the expense of understanding the meaning of mathematical assertions. Thus, a proposition is regarded primarily as a stepping stone to the next proposition. What is seriously lacking is an understanding that a proposition is an efficient way of gathering together a lot of mathematical information. In many cases, the students were able to repeat various mathematical "laws" but were stymied when asked to illustrate them or apply them in a given instance. Furthermore, even the best students in the class had a very weak idea of what constituted a valid mathematical argument and it seemed unwise to push this side of mathematics too far.

As to 2): In close connection with the previous point, it was clear that many of the students did not relate mathematics to any notion of reality. To illustrate, at one juncture, the students were asked to compare the length of two segments that they had randomly drawn themselves. They were asked to compare the lengths experimentally using straight edge and compass. Some students rejected their own findings because the answers did not come out a whole number after three or four bisections. As one student put it "math problems are supposed to come out even". Apparently one reliable way of checking what is drilled into the students in primary and high school is to see if the answer is a single integer. This has had the effect of divorcing mathematics from real life to the extent that the previous astounding quote was possible. We, therefore, strongly felt that a substantial portion of the course should be given in "laboratory" dealing with physical objects.

As to 3): The reason here is two-fold. First of all, since most prospective teachers teaching the course will not be intrinsically motivated to mathematics or be motivated by the applications of mathematics, some external motivation must be supplied. A source of motivation is the possibility of using the material of the course in a future classroom situation. In fact, the greatest show of enthusiasm I saw was when one of the girls was trying out some of the course material in her practice teaching. A second point, of course, is that if a thorough understanding of the material is not achieved by all students, these students will at least have acquired some useful devices for the classroom situation.

✓ The subject matter of the course consisted of a study of the positive real numbers in connection with the measurement process, the study of the whole number line as a one dimensional vector space and the study of vector geometry in the plane. Our reason for this choice of subject consisted, in part, of a desire to counterbalance the recent trend to base all arithmetic on sets, which has had the effect to emphasize the discrete and de-emphasize the continuous and geometric aspects of arithmetic. Due to some debugging of early material, there was not enough time left to adequately treat the vector geometry in the plane. We expect that this material will be covered, tested and revised in the current spring term.

So far, we have developed "laboratory material" such as balances, weights, ruler and compass methods, and a gadget for adding vectors in the plane quickly. A laboratory manual for the first two-thirds of the material has been written. However, it will need to be seriously revised. A final draft of the first portion of the text is currently being written. We anticipate that after further experimentation this term (especially on the last third of the material), we should have a complete package, consisting of text, laboratory manual and laboratory materials.

TABLE OF CONTENTS

Chapter I. Measurement and the Positive Real Numbers

1.1 Introduction. The need for a continuous number system 1

1.2 Inequality 3

1.3 The transitive law for inequalities 6

1.4 Experiments with the transitive law 9

1.5 Equality and equivalence 13

1.6 Equivalence relations 16

1.7 Addition 24

1.8 Multiplication by a positive integer 29

1.9 The Archimedean Principle 32

1.10 Halving objects 36

1.11 Nested intervals 41

1.12 Dyadic expansions 48

1.13 Dyadic expansion of integers 57

1.14 Computation with dyadic expansions 65

1.15 Computation with Infinite Dyadic Expansions 85

Chapter I

Measurement

1-1 Introduction

The purpose of this chapter is to examine the notion of measurement with some care, and also to study the real numbers with emphasis on the role they play in describing the measurement or the size of various objects. It may be remarked, in passing, that by the real numbers we mean all the numbers which are usually used in arithmetic -- that is, all the numbers on the number-line. One of the major objectives of this course is to deepen the student's understanding of the real numbers. This is important mathematically and because of its close connection with the mathematics curriculum of the elementary school.

In recent years, there has been a tendency, at all levels of the educational process, to base arithmetic on the operations of set theory. This has led to heavy emphasis upon the discrete aspects of arithmetic as opposed to the continuous aspects which arise in a natural way from the process of measurement. We prefer to emphasize the continuous, and our procedure will be inductive rather than axiomatic. That is, we will derive various rules or properties of the algebra of measurement as abstractions from experiments or experiences which have geometric or intuitive content.

Our experiments will center on weights and on lengths, although analogous experiments could be applied in any situation where measurable quantities are obtained.

1-2. Inequality

The most primitive notion underlying any situation in which some kind of measurement plays a role is that of inequality. As examples of the type of situation we have in mind, we may list the following: one weight weighs less than another, one stick is shorter than another, or one baseball team is inferior to another. Of course, there are many ways to define what is meant by the statement that team A is worse than team B; one possibility is that team B won the last game they played, another possibility might be that over the full season team B won more games from team A than it lost. The reader should have no difficulty in choosing other possible definitions.

Our first experiments are with the notion of weight. Here, the measurement or more precisely, the comparison, is determined by a balance. Object A is put on one side of the balance, and object B is put on the other side. If the side containing A goes up while the side containing B goes down, we say that object A is lighter than B and write $A < B$. If side A is the one which goes down, we write $B < A$. The sign $<$ is to be read as less than. (It could be that different observers comparing A and B will arrive at different observations. Thus one observer may "see" that A is lower, while the other observer cannot decide which is lower. In our discussion, we shall avoid such problems by assuming that there is an objective reality which is seen by all observers.) In the present context, where the comparison is that of weight, we might use the terminology larger than (and write >) instead of less than. However, in order to maintain mathematical consistency and simplicity we will use only the symbol for less than.

It is extremely important that it be understood that this notion of comparison of weights has nothing to do with numbers. (In particular, such things can be taught to first grade children.) We do not say that object A weighs so many ounces and that object B weighs a certain number of ounces. Our sole assertion is that there is a comparative statement relating objects A and B. At a later stage, we shall analyze the mathematical properties of this relationship; eventually this will allow us to introduce the real numbers as representing the measurement of such things as weights.

We shall also experiment with lengths which are represented concretely by sticks or line segments. Here, we compare sticks A and B by placing one on top of (or against) the other such that both have an end in common -- for example, we might stand them both up on the table. If stick B extends beyond stick A, we say that A is shorter than B (or that B is longer than A). Thus, here too we may write $A < B$ and say that A is less than B. Naturally, we are making the underlying assumption that the result of this comparison is not affected by moving the sticks around in space or by which endpoints we take as common to both. Of course, the same assumption of invariance of comparison under motion applies also the case of weights. Again we emphasize that our comparison has nothing to do with numbers; we do not "measure" each stick -- all we do is compare them.

1-3. Transitive Law for Inequalities

The first fundamental property of our "less than" relationship for weights or lengths is the transitive law. If A is lighter than B and B is lighter than C, then A is lighter than C; in symbols, if $A < B$ and $B < C$ then $A < C$. This rule, which is known as the transitive law, is so obvious that we often take it for granted. It is certainly obvious for the case of weights. It is equally obvious for the case of lengths; that is, if A is shorter than B and B is shorter than C, then A is shorter than C.

On the other hand, the transitive law does not hold in all situations of every day life where we make comparative statements. Consider, for example, the comparison of baseball teams mentioned earlier. If the Minnesota Twins are not as good as the Red Sox (that is, Twins < Red Sox) and also Red Sox < White Sox then it does not follow in practice that Twins < White Sox. The reason for the failure of the transitive law in this context is that more than one factor enters into the winning of a single ball game or series of ball games and these factors may not combine. It might be that Minnesota hitters hit White Sox pitching very well but do poorly against the Red Sox. By the same token the Twins may have pitchers who lose consistently to the Red Sox (because of the special dimensions of Fenway Park) but who specialize in certain pitches that the White Sox hit very well. In addition, the Red Sox may be weak in fielding which causes them to lose to the White Sox. Thus the factors that determine who wins or loses may not combine, and it is quite possible that Twins $<$ Red Sox, Red Sox $<$ White Sox, and White Sox $<$ Twins.

An example which illustrates this point is the children's game commonly known as "Rock, Scissors and Paper". The rules of the game are as follows: There are two players; each places one hand behind his back; then the hidden hands are brought forth simultaneously. Each child displays either a clenched fist representing a rock, or his open hand representing paper, or two fingers representing a pair of scissors. If one child displays a fist and the other an open hand, then the one with open hand wins because paper wraps rock. Furthermore, scissors wins over paper because scissors cuts paper, and rock wins over scissors because rock can break scissors. In short, the rules of the game are, rock $<$ paper, paper $<$ scissors, scissors $<$ rock, which means that there is a clear-cut violation of the transitive law. By stretching things a bit, one might say that the transitive law breaks down precisely because the relationship between rock and scissors is entirely different from the relation between scissors and paper...Roughly speaking, we may say that the transitive law holds when our comparison is based on a single simple quantity such as weight or length, rather than on some complex combination of factors.

Eventually, we intend to express the relationship of inequality (that is, less than) in terms of numbers. More precisely, one of our goals is to assign numbers to objects in such a way that relations between objects are reflected by corresponding relations between the assigned numbers. In other words, if A and B are objects and "less than" compares them in weight or length we would like to assign numbers to A and B such that $A < B$ if and only if the number associated with A is less than the number associated with B. Thus, the relation of < for objects will correspond to the relation of < for the associated numbers. Since, as is well known, the transitive law holds for numbers, it becomes absolutely essential that (in order to preserve the < relation under our correspondence) we deal with objects for which the < relation is transitive. We also want to emphasize that unlike the example of the baseball teams, mathematicians use the symbol < only in cases where the transitive law holds. This accounts, in part, for our emphasis on weights and lengths.

Topics for Discussion:

1. How would you undertake to teach small children about the transitive law?
2. What is the meaning of the phrase "if and only if?" Are you acquainted with other ways of saying the same thing? What is a "necessary condition"? What is a "sufficient condition"? What is meant by a "necessary and sufficient" condition? What is a converse?

1-4. Experiments with the Transitive Law

Our first experiments concerning the transitive law for weights are based upon use of the balance. If the balance tells us that object A weighs less than object B and also that object B weighs less than object C, then one observes, experimentally, that when A and C are compared it turns out that A is less than C. By performing this experiment several times with different choices of A, B, and C we may satisfy (i.e. convince) ourselves experimentally that the transitive law does, in fact, hold for weights. Once this stage has been reached, the transitive law may be used as a principle of deduction. Thus if $A < B$ and $B < C$ then we may conclude that $A < C$ without making use of the balance. Furthermore if, in addition, we know that $C < D$ then the transitive law enables us to deduce that $B < D$ and $A < D$. It is clear then that the transitive law may be used to "telescope" a series of inequalities -- for example, if we have also, $D < E$, $E < F$, $F < G$, $G < H$ then one conclusion is $A < H$.

In this connection, an interesting experiment is to start with a reasonably light weight A and have each of the students construct, in succession, a heavier weight. In other words, the first student with his balance constructs weight B slightly heavier than A, then the second student uses his balance to construct weight C slightly heavier than B, and this process continues as the weights are passed around the room. If one stops this experiment at any point and compares the end weight with the original weight A, then the result is always A $<$ end weight. Thus, the transitive law and its consequences hold experimentally without any difficulty. On the other hand, we shall see later that difficulties arise when one tries to deal experimentally with the transitive law for equalities.

The transitive law can be conveyed effectively to children as a matter of organizational efficiency. Suppose they are given a large number of objects and asked to record all the comparative statements that can be made relating any pair of these objects. For example, suppose that each child is given a weight (clearly, lengths could be used instead of weights). The teacher selects pairs of children (many of them) and asks them to compare their weights. The children should soon observe that the most efficient way to organize all this information is to order all the objects according to increasing weight -- for this enables them to deduce the relation between any pair of objects from this ordering and the transitive law. Of course, heavy use is made of the transitive law in ordering all the objects according to increasing weight.

A useful pedagogical device which may be introduced at this point, with the purpose of hammering home the use of the transitive law, is to play a guessing game with the following rules. A certain number, call it n, of objects are given and arranged according to weight -- for convenience, we may write $A_1 < A_2 < A_3 < \dots < A_n$. Someone selects one of these n objects, and the others must then guess which object was chosen. The guessers are permitted to ask questions of a single type (which questions do not count as guessers), namely -- is the unknown object greater than (or less than) the i th object A_i. Naturally, guessers may be made even before any questions are asked, but after a few trials the children may get some feeling as to how to ask questions efficiently, so that after as few questions as possible they have no doubt as to which is the unknown object. It is easy to see that for n=3 objects, 2 questions suffice for determining the unknown object with certainty -- while 1 question does not suffice. For any n, let q denote the minimal number of questions after which we can pick out the unknown object with certainty. We have noted already that if n=3 then q=2. It may then be observed that if n=4 then q=2, while if n=5 then q=3. Continuing our systematic examination of the connection between n and q, we see that if n=8 then q=3. It follows then that for for n=6 or 7 we have q=3. The next case to consider is $n=16=2^4$-- and by now it is fairly clear that q=4. As before, it follows that if $n=9$, 10, 11, 12, 13, 14 or 15 then also $q=4$. What about the general rule? This is essentially within our grasp. If n is a power of 2, say $n=2^k$, then by simply extending the procedures used before, we see that $q=k$. If n is not a power of 2, then n lies between two consecutive powers of 2 -- that is, $2^{k-1} < n < 2^k$ -- and we get $q=k$. For example, if $n=100$ then $2^6=64 < n=100 < 2^7=128$ so that $q=7$.

Of course, one does not discuss n and q explicitly or in a formal sense with children -- one simply does many examples, and leads them to discover the pattern. Such a line of exploration introduces children to some uses of powers of 2, and serves as preparation for the eventual study of binary expansions.

1-5. Equality and its Properties

If objects A and B are placed on different sides of a balance and neither side goes down, that is, if the two sides balance, then we say that object A is equal in weight to object B. For simplicity, we then write A=B, although such a notation obviously leaves much to be desired. In similar fashion, segment A is said to be equal in length to segment B if neither one is longer than the other. We shall deal with weights, although the same sort of discussion would apply equally well to lengths.

The first fundamental observation about the relationship of equality is again the validity of the transitive law. That is, if A=B and B=C then A=C. However, in contrast to the transitive law for inequality, the transitive law for equality is frequently an idealization from experience rather than something that always holds true in practice. Thus, if we have objects A, B, C, D, E with A=B, B=C, C=D and D=E then standard rules of reasoning lead to the conclusion that A=E. Unfortunately, the experiment corresponding to this assertion often breaks down. In fact, suppose that one student starts with object A and produces object B of equal weight. He keeps object A and gives B to the next student who constructs object C equal in weight to B. He then gives C to another student and the same process is repeated; this goes on as many times as desired -- 10 will usually suffice, but one may prefer to have every student participate. If the last object constructed is compared with A, they frequently turn out to be of unequal weight.

The reason for this apparent contradiction of the rules of logic is, of course, the inaccuracy of our balance. There is a certain amount of experimental error involved; thus although A and B balance on our rough balance, they are probably not really equal in weight, and the use of a more delicate and accurate balance could show this. Now, such errors can accumulate sufficiently so that they do indeed show up even on our rough balance; this is why the experiment led to an unexpected result. Unfortunately, this accumulation of error is unavoidable. If we were to use extremely delicate balances, the same trouble would arise, because, after all, no balance is truly perfect.

It may be remarked that if this experiment is repeated a number of times, it will turn out that sometimes the end product is lighter than A, sometimes it equals A, and sometimes it is heavier than A. If things work reasonably well, the end product turns out to be lighter than A or heavier than A with equal frequency. This indicates that the break-down of the transitive law for equality does not reflect something that is fundamentally missing from the relation -- rather, it is due simply to accumulation of experimental error. The cases in which the end product is equal to A in weight occur precisely when the various experimental errors cancel each other -- some students may produce weights which are too heavy while others may produce weights which are too light.

In summary, the transitive law for equality is a rule which we regard as holding in an ideal situation. According to our viewpoint, the equality represented by a balance is merely a crude approximation to the ideal equality that we would expect to hold for an ideal balance.

Question: If A=B and B $<$ C, what if the relation between A and C ? Explain.

Discuss whether this is an experimental fact or a law of logic.

1-6. Equivalence Relations

If we have objects A and B and are dealing with weights (analogous remarks will apply to lengths) then our previous notation involves writing $A=B$ to signify that A and B are equal in weight. This notation has an unfortunate aspect which could conceivably lead to confusion, for it is customary to interpret a statement like $A=B$ in terms of "being identical" -- that is, object A is the same as object B, so that A and B are possibly different names for the same object. In the interest of precision we shall temporarily use the notation $A \overset{w}{=} B$ to mean that A and B are equal in weight; another possible notation would be $w(A)=w(B)$ (We shall revert to our old notation after this section.)

This definition of equality in weight, $A \overset{w}{=} B$ or $w(A)=w(B)$ implies that we are focusing attention only on what the balance tells us. Thus, for our purposes, a cup of coffee and a soggy doughnut are the "same" if they balance. The important point is that this relation $\overset{w}{=}$ of balancing allows us to introduce an abstract notion called "weight" to each real object: $w(A)$ is the weight of the object A. We mean this is in the same sense that we attach the color green to all green objects. We may then consider the idea of "green" as an abstract notion in its own right. Note that, as yet, we have no right to consider weight as a number any more than we can consider color as a number.

In the English language we tend to distinguish between adjectives and nouns. In a certain sense this distinction is artificial and purely a matter of usage or convenience. We are not accustomed to saying "a green" when we mean any green object or "a fat" when we mean any fat person. (There are exceptions; we do say "a square" to mean any square figure.) In mathematical discourse, however, it is quite common to drop the distinction between adjective (or other modifying word) and noun. We talk, therefore, about "the weight A" when we really mean "any object whose weight is w(A)". We proceed to analyze this idea of introducing an "abstract" notion such as weight.

The general mathematical setting in which the preceding notion of equality in weight (or of equality in length) should be viewed involves the concept of an equivalence relation. We now proceed to explain what is meant by an equivalence relation in a somewhat abstract setting. Consider an arbitrary set S whose elements or numbers are denoted by A, B, C, D,...and such that there is given some relation, denoted by R which may or may not hold between any two elements of S. Thus, for any given pair (A,B), in the given order, with $A, B \in S$ we write A R B when A is related to B (that is, when A and B satisfy the relation) and $A \overset{x}{R} B$ when A is not related to B. Some concrete examples should prove helpful at this point:

1) Let S be the set of all integers, that is, $S = \{\ldots, -2, -1, 0, 1, 2, 3, \ldots\}$ and let the relation R be "less than" (in symbols, $<$). Then $A \; R \; B$ means $A < B$, while $A \overset{x}{R} B$ means that A is not less than B (that is, $A \geqslant B$).

2) $S = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and R is $\leqslant$ (less than or equal to).

3) S is the set of all objects and R is the relationship of being equal in weight. Here, for $A, B \in S$, $A \; R \; B$ means that $A \overset{w}{=} B$. Of course, this may also be done for lengths.

4) S is the set of all objects (i.e. weights) and R is the relation of less than. Here, $A \; R \; B$ means that A is less than B in weight. In our old notation this would be expressed as $A < B$; however, in keeping with our remarks at the beginning of this section, it might be preferable to write $A \overset{w}{<} B$ or $w(A) < w(B)$.

5) S is the set of all real numbers and R is $\leqslant$ -- so $A \; R \; B$ if and only if $A \leqslant B$.

6) S is the set of all real numbers and R is $=$.

7) S is the set of all triangles in the plane, R is the relation "has the same area as" -- so $A \; R \; B$ signifies that A has the same area as B. Among other possible relations on this same set S we may mention "is congruent to", "is similar to" or "has the same perimeter as".

8) S is the surface of the earth, and A R B means that A has the same latitude as B.

9) $S = \{0, \pm 1, \pm 2, \ldots\}$ is the set of all integers and A R B is taken to mean that A-B is divisible by 7.

Returning now to an arbitrary set S with a relation R on it, if the following three properties are satisfied we say that R is an equivalence relation.

(I) A R A for all $A \in S$ (reflexive law)

(II) If A R B then B R A for all $A, B \in S$ (symmetric law)

(III) If A R B and B R C then A R C for all $A, B, C \in S$ (transitive law)

The reader may verify easily that the reflexive law is satisfied in examples 2, 3, 5, 6, 7, 8, 9, and that it is not satisfied for examples 1 and 4. Note that the reflexive law in example 3 is really a logical "fiction"; it cannot be verified experimentally with a balance because the object A cannot be placed on both sides of the balance simultaneously. There is only one object A, and any copy of it is obviously not the same as object A. Thus, for our ideal balance, we are really making the assumption that if the same object could be placed on both sides of the balance simultaneously then both sides would balance--i.e. that an object weighs the same as itself.

The reader may also verify that the symmetric law holds for examples 3, 6, 7, 8, 9 and that it does not hold (which means that we need produce only one case in which it breaks down) for examples 1, 2, 4, 5. Note that in example 3 the symmetric law reflects the underlying assumption that our ideal balance is not "biased"; in other words, if A and B balance when A is placed on, say, the left side of the balance and B is on the right side (i.e. if A R B) then they also balance when A is on the right side and B is on the left (i.e. B R A). In particular, writing A R B involves distinguishing one side of the balance.

Finally it is easy to see that the transitive law is satisfied for all the examples 1 through 9.

Problem: Define a set S with a relation R such that

- a) R is symmetric and transitive but not reflexive
- b) R is reflexive and transitive but not symmetric
- c) R is reflexive and symmetric but not transitive
- d) R satisfies only the reflexive law
- e) R satisfies only the symmetric law
- f) R satisfies only the transitive law

Examples 3, 6, 7, 8, 9 are, all of them, equivalence relations, and the reader may easily produce other examples of equivalence relations. The mathematical importance of the notion of equivalence relation is that, in such a situation, the set S can be partitioned into disjoint subsets (which subsets are usually known as "equivalence classes"). In detail: we say that A is equivalent to B (with respect to R, of course) when A R B, and then for every $X \in S$ we let $[X]$ denote the set of all elements of S which are equivalent to X -- symbolically, $[X] = \{A \in S \mid A\ R\ X\}$. A subset of form $[X]$ for $X \in S$ is known as an equivalence class -- or the equivalence class determined by X. The fundamental properties of these equivalence classes are as follows:

i) $A \in [A]$ for any $A \in S$; in words, each element of S belongs to the equivalence class which it determines.

ii) If $B \in [A]$ then $[B] = [A]$; in words, any element of an equivalence class determines the class.

iii) $[A] = [B] \Longleftrightarrow A\ R\ B$; in words, two elements of S determine the same equivalence class if and only if they are equivalent.

iv) If $[A] \cap [B] \neq \emptyset$ then $[A] = [B]$;in words, if two equivalence classes have an element in common then they are identical.

v) $[A] \cap [B] = \emptyset \Longleftrightarrow A\ \overset{X}{R}\ B$

As for the proof of these properties, i) is immediate in virtue of the reflexive law. To prove ii), note that, using the symmetric law, $B \in [A] \Longleftrightarrow B\ R\ A \Longleftrightarrow A\ R\ B \Longleftrightarrow A \in [B]$. Then $X \in [B] \Longrightarrow X\ R\ B \Longrightarrow X\ R\ A$ (since B R A)$\Longrightarrow X \in [A]$. This means that $[B] \subset [A]$, and in similar fashion (that is, by a symmetrical argument) $[A] \subset [B]$-- therefore, $[A] = [B]$. The proof of iii) is now simple: $[A] = [B] \Longrightarrow B \in [A] \Longrightarrow B\ R\ A \Longrightarrow A\ R\ B$, and $A\ R\ B \Longrightarrow A \in [B] \Longrightarrow [A] = [B]$. To prove iv), observe that if $C \in [A] \cap [B]$ then $C \in [A]$ and $C \in [B]$, so $[A] = [C] = [B]$. The proof of v) is left to the reader.

From what has gone before we see that if S is a set on which we have an equivalence relation then two equivalence classes are either disjoint (that is, having no element in common) or identical -- not both -- so that S breaks up into disjoint equivalence classes. Thus we can form a new set, denoted by S/R, whose elements are the distinct equivalence classes $[A]$, $[B]$, etc...

If we are considering the set S of all material objects with the equivalence relation R of "is equal in weight to", then each equivalence class consists of all objects which happen to have the same weight. Mathematically, we may then think of each such class as a new object in its own right. In this way, each equivalence class has a weight associated with it (one might even go further and say that each equivalence is a weight). This is entirely analogous to considering the set of all colored objects with the relation "is the same color as". The equivalence classes consist of all objects having the same color -- so each equivalence class may, for all practical purposes, be considered as the color itself, and the set S/R here is just the set of all colors.

Returning to weights, we observe that for the equivalence classes we have a natural notion of $[A] < [B]$ -- namely, when $A \overset{w}{<} B$. A key point here is that $[A] < [B]$ is "well defined"; this means that the definition does not depend on the choice of representatives for the equivalence classes. In other words, if $[A'] = [A]$ and $[B'] = [B]$ then (see the question at the end of sec. 1-5) $A \overset{w}{<} B \Longleftrightarrow A' \overset{w}{<} B'$. Consequently, the notion of less than can be regarded not only as a relation between objects but also as a relation between weights -- that is, as a relation between equivalence classes.

Problem: Discuss several equivalence relations (especially example 9) and describe the equivalence classes.

1-7. Addition and its Properties

For convenience, we shall deal in this section only with weights, and leave it to the reader to consider the analogous situation of lengths. Our purpose is to show that we can combine weights in such a manner that the usual rules for addition hold.

Consider any two objects A and B, and combine them by lumping them together into a single pile. This pile may be viewed as a new object which we denote by A+B. From the point of view of our balance, A+B means simply that both A and B are placed together on the same side of the balance. Since it clearly does not matter in what order A and B are placed on the same side of the balance, there is no way to distinguish between A+B and B+A; therefore, we must view A+B and B+A as the same object -- that is, A+B = B+A.

Suppose we now take additional objects A' and B' with $A' \overset{w}{=} A$ and $B' \overset{w}{=} B$. Then we may form A'+B' and verify experimentally that $A'+B' \overset{w}{=} A+B$. (Thus we are verifying here the familiar phrase: adding equals to equals gives equals.) By the definition of equivalence classes for objects with respect to the relation of "equal in weight" we know that $A' \overset{w}{=} A$ means that $A' \in [A]$ and $B' \overset{w}{=} B$ means that $B' \in [B]$. Our experiment therefore tells us that if $A' \in [A]$ and $B' \in [B]$ then $A'+B' \in [A+B]$. This says that we can define the notion of addition on the set of all equivalence classes -- that is, on the set of weights. More precisely, if we are given two weights (that is, equivalence classes) $[A]$ and $[B]$ then we define

$$[A] + [B] = [A+B]$$

This operation seems to depend on the choice of the objects A and B, but the thrust of our experiment is that this operation of addition of weights is well-defined -- in other words, if A' and B' are equal in weight to A and B respectively then A'+B' is equal in weight to A+B; in symbols, $[A'] = [A]$ and $[B'] = [B]$ together imply $[A'+B'] = [A+B]$. To put it still another way, if A' and A belong to the same equivalence class and also B' and B belong to the same equivalence class then A'+B' and A+B belong to the same equivalence class. In short, the addition of weights does not depend on the choice of objects of the given weights.

This operation of addition provides a crucial step towards our goal of assigning numbers to abstract properties such as weights. With this objective in mind we need, first of all, to observe that the usual rules for addition of numbers are valid for this operation of addition of weights. We also need to understand how this relation of addition interacts with the relation of inequality (i.e. less than).

Let us sketch briefly several experiments with weights which lead to important properties of addition. Having been quite careful heretofore in distinguishing between an object A and its weight $[A]$, we shall now find it convenient to drop this distinction. This should cause no difficulty, as it should be clear from the context what is meant.

1) Given three weights A, B, and C we may construct the weights $D = A+B$ and $E = B+C$. As indicated earlier, it is clear that $A+B = B+A$; that is, of course, known as the commutative law for addition. We may also show experimentally that $D+C = A+E$. This assertion is usually written as

$$(A+B) + C = A + (B+C)$$

for all weights A, B, C and is known as the associative law for addition. Notice that this implies, for example, that

$$\big((A+B) + C \quad + F = (A+B) + (C+F) = A + \big(B + (C+F)\big) \text{, etc...}$$

In short, in order to add several weights it doesn't matter where the parentheses are placed -- that is, in what order the additions are performed. Furthermore, the end result is the same as would be obtained by simply putting all the weights in the same pan of the balance. Thus, there is no ambiguity about the meaning of an expression of form $A+B+C+F+G$; even more, in virtue of the commutative law, this is equal to C+G+B+E+A or to any other sum of the same weights in whatever order.

2) Suppose that A and B are weights with $A < B$ then, as noted in section 1-6, if $A' = A$ and $B' = B$ we may verify experimentally that $A' < B'$. It is also equally easy to check that for any weight C we have $A+C < B+C$. If, in addition, $C < D$ then it is a consequence of the transitive law that $A+C < B+D$. Of course, this can also be checked experimentally.

3) Suppose we have weights A, B, C, D, X, Y with $A < X < B$ and $C < Y < D$ then it follows from what has gone before that $A+C < X+Y < B+D$. In other words, the weight X+Y is boxed in between A+C and B+D. It should further be noted that this rule, which applies for lengths also, involves a certain loss of information. We may illustrate what is meant by examining an analogous situation.

Suppose we are dealing with real numbers. As is usually taught in grade school we say, for example, that $x = 5$ to the nearest integer when $4\frac{1}{2} \leq x < 5\frac{1}{2}$. Suppose further that $7\frac{1}{2} \leq y < 8\frac{1}{2}$, that is, $y = 8$ to the nearest integer. Therefore, adding inequalities, we have $12 \leq x + y < 14$ and we can no longer say what $x + y$ is to the nearest integer -- it could be 12, 13 or 14 depending on appropriate choices for x and y, just so long as they are within the prescribed bounds. Thus, the addition of inequalities has involved a loss of information -- that is, when dealing with the notion of "to the nearest integer" addition is not determined to the nearest integer.

Problems:

1) Discuss the transitive law for weights -- that is, for equivalence classes -- and its experimental verification.

2) How are the possible experimental errors in this section related to the desired theoretical statements? What should children be told about experimental errors?

3) As in example 9 of section 1-6, let $S = \{0, \pm 1, \pm 2, \ldots\}$ be the set of all integers, and let R be the relation such that A R B means that A-B is divisible by 7. Show that R is an equivalence relation, and describe the equivalence classes. Define addition on the set of equivalence classes. Define, if you can, an order (that is, a relation of less than) on the set of equivalence classes; is it transitive? does it satisfy the condition that $[2] < [B]$ implies $[2+8] < [B+8]$ where $[2]$, $[B]$, $[8]$ are equivalence classes? Can you generalize this entire example?

1-8. Multiplication by a Positive Integer

For convenience we shall continue to deal with weights, and to use the symbols A, B,... to denote both an object and its weight.

From the preceding section, we know how to add weights; thus, for any weight A we may define 2A = A+A, 3A = A+A+A, and, in general, for any positive integer n, nA = A+A+....+A, where there are n copies of A in the sum on the right. Note that for n=1, ~~the definition says that~~ $_1A = A$. This operation, in which we take a positive integer and a weight and "combine" them to get a weight may be called "multiplication by a positive integer"; it will be generalized significantly later.

There are several natural and important properties of this operation. From the associative law for addition it follows that if m and n are positive integers and A is an arbitrary weight then

$$(m+n)\,A = mA + nA$$

for instance when $m = 2$ and $n = 3$ $(2+3)\,A = 2A+3A$

and

$$(mn)\,A = m\,(nA)$$

so $6A = 2(3A)$

Note that in the first of these equations the addition on the left side is for integers, while on the right side it is addition of weights. In addition it follows from the associative and commutative laws for addition that if n is any positive integer and A and B are arbitrary weights, then

$$n\ (A+B) = nA + nB$$

Let us illustrate the steps of the proof for the case $n = 2$;

$$2(A+B) = (A+B) + (A+B) = \big((A+B) + A\big) + B = \big(A+(A+B)\big) + B =$$
$$\big((A+A) + B\big) + B = (A+A) + (B+B) = 2A + 2B.$$

Despite the fact that the distributive laws $(m+n)A = mA + nA$, and $n(A+B) = nA + nB$ and the "associativity" property $(mn)A = m(nA)$ are logical consequences of the rules for addition, it is of some value to verify them experimentally. When this is done, even with a great deal of care, the experiment may fail -- these laws are really idealized statements, and they are more than mere tautologies.

It is clear that $m = n$ implies $mA = nA$ for any A. Conversely, we observe, that for any A, if $mA = nA$ then $m = n$. Of course, this too is an idealized statement; in fact, if A is sufficiently light our imprecise balance may even be unable to distinguish between A and 2A -- in other words, the balance would say, $A = 2A$. It should also be pointed out that this rule (i.e. cancellation law) is not obvious to young children. As a matter of fact, the simpler notion that counting a set of discrete objects always yields the same number is something of which they are not certain. This explains, in part, why they will often count the elements of a set in several ways.

Finally, we may observe (either experimentally or as a consequence of properties of addition) that multiplication by a positive integer preserves the relation of less than -- in other words, if $A < B$ then $nA < nB$ for any positive n.

Problem: If $m < n$ what relation exists between mA and nA ? Explain.

Discuss whether this is an experimental fact or a law of logic.

1-9. The Archimedean Principle

Once we have introduced the notion of multiplication by positive integers we can begin to make some refinements on the relation of inequality. We have seen that for all positive n, $nA < nB$ is a consequence of $A < B$ and it is instructive to discover statements relating a multiple of A with some other multiple of B (that is, comparing nA and mB) that are not consequences of $A < B$. For instance, suppose that $A < B$; we may then ask, how does 2A compare with B. If $B < 2A$ then the pair of inequalities $A < B < 2A$ surely provide more information than the single relation $A < B$. If on the other hand $2A < B$, we might then compare 3A and B, and get perhaps $3A < B$ -- or going one step further, perhaps $3A < B < 4A$. The question we are really considering here is the following: there are two sets of inequalities $B < 2B < 3B < \ldots\ldots < mB < \ldots\ldots$ and $A < 2A < 3A < \ldots\ldots < nA < \ldots\ldots$ and our problem is how to interleave these two sequences -- that is where to place the multiples of A in the sequence of multiples of B.

In order for this procedure to be effective, we would certainly want to know that A and B are comparable (in magnitude)-- in other words, that if $A < B$ we do not have all multiples of A less than B. What we really want then is that, for any A and B there exists an integer n, which may be very large, such that

$nA > B$ (of course, the roles of A and B here are interchangeable). When this property does hold, it is known as the Archimedean Principle.

The Archimedean Principle is easy to verify experimentally in the case of weights or of lengths. For example, if A is a drop of water and B is a house then taking enough drops of water (that is, taking n sufficiently large) we get $nA > B$. On the other hand, the Archimedean Principle need not hold in all situations where there is an interplay between addition and inequality. Let us give an example.

Consider all possible words that can be formed from the 26 letters of the English alphabet, where by a word we mean any finite sequence of letters. Thus abcdef is a word, as are cat and dog. We may then form a nonsense dictionary of all such words, where the words are placed in lexicographic order -- that is, the usual dictionary order. This provides us with a notion of less than; for example $\text{aa} < \text{aba} < \text{abcdef} < \text{cat} < \text{cow} < \text{dog} < \text{teacher} < \text{xerox}$. Of course, this relation of $<$ is transitive. Now, let us define addition of words simply as juxtaposition -- for example, cat + dog = catdog and abcdef + bcxy + adcdefbcxy. In terms of this addition if $a < b$ then a "plus" $c < b$ "plus" c, that is $ac < bc$. The reader will notice that if A and B are words with $A < B$ then $nA < B$ for all n. Thus, the Archimedean Principle is violated.

Problem: In formulating the Archimedean Principle it is important that we keep adding A to itself. If it is not always the same A that is added, then the principle need not hold. For instance, for the real numbers we have Zenos' paradox, which says essentially that if we add $a_1+a_2 + a_3 + \ldots\ldots + a_n + \ldots\ldots$ then we may not be able to add enough terms to get a result exceeding any fixed b.

a) Can we add enough terms of $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots\ldots + \frac{1}{2^n} + \ldots\ldots$ to get an arbitrarily large number?

b) Can we add enough terms of $1 + 1/2 + 1/3 + 1/4 + \ldots\ldots + 1/n + \ldots\ldots$ to get an arbitrarily large number.

Problem: In the nonsense dictionary example of a non-archimedian system the law for addition is not commutative. Thus, cat + dog = catdog and this is not the same as dog + cat = dogcat. We can improve on our example to make it commutative. Do this by considering a dictionary of complete nonsense where the only words allowed are in alphabetical order. Thus, cat or dog would not be allowed but act and dgo would. Now define addition as juxtaposition followed by arranging the new word in alphabetical order. Thus

$$act + dgo = acdgot$$

Show that all the rules we have described so far are satisfied except the Archimedean Principle.

Topic for Discussion:

Can you think of human value judgements where the Archimedean Principle is violated?

For instance compare human life with money: one human life is worth more than any amount of money.

1-10. Halving of Objects

In the next section, we shall combine the standard properties already at our disposal with the Archimedean Law in order to derive and organize more precise information connecting given objects A and B than we have been able to get heretofore. To do this, we need to make one simple and rather natural physical assumption -- given any object A there exists (and presumably, we can find) an object B such that $2B = A$ -- or equivalently, we may write $B = \frac{1}{2}A$. If B' is any other object such that $2B' = A$ then $B' = B$.

In this section, we shall discuss some of the elementary properties of this process of halving. First of all, it should be noted that the mechanics of carrying out such a division into two equal parts experimentally can lead to all kinds of technical difficulties. For example, if object A is a weight consisting of a container of water then, even in this simple case, it takes time (and usually several approximations) to get $\frac{1}{2}A$. However, for lengths as represented by segments there is a well-known mechanical procedure of dividing a segment in half by use of ruler and compass. Because of this, our discussion will center on segments; another possible advantage in dealing with segments is that visual intuition may be helpful in understanding what is going on.

It should be noted that the choice of division by 2 is little more than a matter of taste and convenience. We could equally well deal theoretically with the division of an object into n equal parts, for any integer $n \geq 2$. (As shall become clear later, our usual number system is based on the case $n = 10$.) Of course, it is more difficult to divide an object into 3 or more equal parts than to divide it in half -- so that physical convenience or efficiency points toward the choice of $n = 2$. As our discussion proceeds it will also be seen that the choice of $n = 2$ leads to some logical and computational advantages.

Let us turn to some of the consequences of "division by 2". Suppose we have two objects A and B; we may take $\frac{1}{2}A$, $\frac{1}{2}B$ and also $\frac{1}{2}(A+B)$, and then observe that

$$\tfrac{1}{2}(A+B) = \tfrac{1}{2}A + \tfrac{1}{2}B$$

To prove this, it suffices, by the definition of halving, to observe that $2(\frac{1}{2}A + \frac{1}{2}B) = \frac{1}{2}A + \frac{1}{2}B + \frac{1}{2}A + \frac{1}{2}B = \frac{1}{2}A + \frac{1}{2}A + \frac{1}{2}B + \frac{1}{2}B = A + B$.

Another useful property of halving is that it behaves correctly for inequalities -- more precisely,

$$A < B \Longleftrightarrow \tfrac{1}{2}A < \tfrac{1}{2}B$$

To see this, we note first that if $\frac{1}{2}A < \frac{1}{2}B$, then according to the rules for adding inequalities, $A = \frac{1}{2}A + \frac{1}{2}A < \frac{1}{2}B + \frac{1}{2}B = B$. In the same way, $\frac{1}{2}B < \frac{1}{2}A$ implies $B < A$; and according to the rules for adding equalities $\frac{1}{2}A = \frac{1}{2}B$ implies $A = B$. Since exactly one of $\frac{1}{2}A < \frac{1}{2}B$, $\frac{1}{2}A = \frac{1}{2}B$, $\frac{1}{2}B < \frac{1}{2}A$ holds, it follows that $A < B \Longrightarrow \frac{1}{2}A < \frac{1}{2}B$ -- thus completing the proof.

Starting from any segments A we have postulated the existence of a segment $\frac{1}{2}A$ for which $\frac{1}{2}A + \frac{1}{2}A = A$ (that is, $2(\frac{1}{2}A) = A$) and noted how $\frac{1}{2}A$ may be constructed with ruler and compass. (Naturally, it is implicit here that A is not too big or too small to be handled with our given ruler and compass.) In general, if n is any integer > 1 and C is a segment such that $nC = A$ then we may introduce a new symbol $\frac{1}{n}A$ for C, and note that $n\ (\frac{1}{n}A) = A$. Perhaps, it needs to be emphasized that although we know how to construct $\frac{1}{n}A$ with a ruler and compass, for any $n > 1$, we are not even assuming at this stage that $\frac{1}{n}A$ exists; our only assumption is that $\frac{1}{2}A$ exists. Now, as observed at the end of section 1-8, $A < 2A$, and consequently $A > \frac{1}{2}A$. If we divide in half again, the result is $C = \frac{1}{2}(\frac{1}{2}A)$; and since $2C = \frac{1}{2}A$ and $4C = A$ we have $C = \frac{1}{4}A$ -- in other words, any A can be divided into four equal parts which are denoted by $\frac{1}{2}(\frac{1}{2}A) = \frac{1}{4}A = \frac{1}{2^2}A$, and such that $A > \frac{1}{2}A > \frac{1}{2^2}A$. Dividing by 2 once more, we see that $\frac{1}{8}A = \frac{1}{2}(\frac{1}{4}A) = \frac{1}{2}(\frac{1}{2^2}A) = \frac{1}{2^3}A$ exists and $A > \frac{1}{2}A > \frac{1}{2^2}A > \frac{1}{2^3}A$. This process of halving may be repeated; we then have the segments

$$\frac{1}{2}\left(\frac{1}{2^r}A\right) = \frac{1}{2^{r+1}}A \qquad r = 1, 2, 3\ldots$$

and

$$A > \frac{1}{2}A > \frac{1}{2^2}A > \ldots\ldots > \frac{1}{2^r}A > \frac{1}{2^{r+1}}A > \ldots\ldots$$

Moreover, according to our basic assumption, this shrinking sequence of segments never stops -- that is, at any stage, it is theoretically possible to divide the segment at hand in half. In practice, of course, the segments we deal with get quite small rather quickly -- for example, if segment A is one mile long and we divide in half 15 times, the result is $\frac{1}{2^{15}}A$ whose length is less than 2 inches. Thus, because our tools are so rough, after a few divisions the segments become too small for physical manipulation ... but the theoretical story continues. In particular, supposing that we cannot physically construct half of a segment of length $\frac{1}{16}$ of an inch and assuming (as is quite reasonable) that the original segment A has length $\leq$ one foot, it then takes no more than 8 divisions by 2 to arrive at a break-down situation where we can no longer divide by 2.

We have already observed that the segments in the sequence $A > \frac{1}{2}A > \frac{1}{2^2}A > \ldots\ldots > \frac{1}{2^n}A > \ldots\ldots$ get small very rapidly. It is also worth noting that they get "arbitrarily small" -- that is, as small as we like, or as close to 0 as we like. More precisely the assertion is that given any segment C, there exists an integer n such that $(\frac{1}{2^n})A < C$ -- so that $C > \frac{1}{2^n}A > \frac{1}{2^{n+1}}A > \ldots\ldots$ The proof of this assertion is not hard. According to the

Archimedean principle there exists an integer n such that $nC > A$. But there always exists a power of 2 which is greater than n -- in fact, the reader may show, by induction, that $2^n > n$. We have then $2^nC > nC > A$, from which it follows that $\frac{1}{2}^n A < C$.

Topic for Discussion:

What is Mathematical Induction?

What is the Binomial Formula?

The reader may show that by either of the above methods $2^n > n$.

1-11: Nested Intervals

In this section we fix a segment A and use it as a standard against which to measure any segment B -- that is, we consider A as a "unit" of measure. It is left to the reader to convince himself at every stage that the discussion carries over to weights.

Given the fixed segment A and any segment B we may compare them in the usual way by placing them one on the other with one endpoint in common. For convenience, let us set things up in such a way that the common end-point is on the left. Along the line determined by the segment A we may consider the segments $A < 2A < 3A < 4A < \dots < mA < \dots$ all of which have a common end-point on the left. According to the Archimedean principle there exists a positive integer m such that $B < mA$ -- in other words, by adding A to itself enough times we get a segment bigger than B. Let nA (n an integer) be the last segment for which $nA \leq B$ -- this implies that the next one $(n+1)A$ is not $\leq B$ -- so

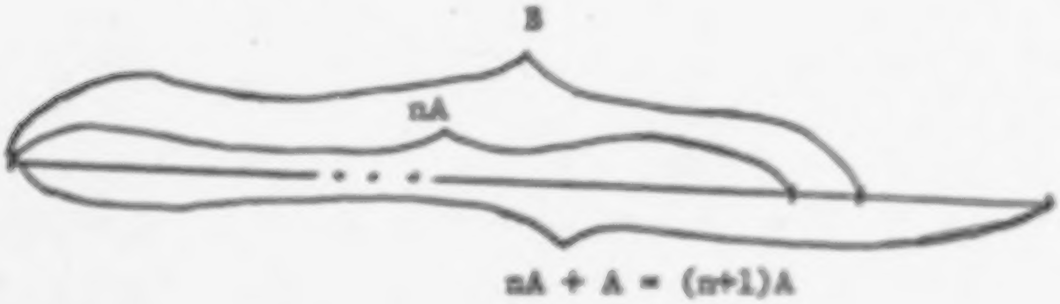
$$nA \leq B < (n+1)A \qquad (*)$$

Note that we may well have $B < A$, so that we must allow the case $n = 0$, too; here we write $oA \leq B < 1A$, where oA may be considered as a formal symbol (with $oA \leq$ any segment, and this implies

$oA <$ any segment) which is introduced in order to permit uniform notation in (*) for all $n \geq o$.

For convenience, when $nA \leq B < (n+1)A$ we shall say that B falls in the "interval" $[nA, (n+1)A)$ and write $B \in [nA, (n+1)A)$. The distinction between the square bracket on the left and the ordinary parenthesis on the right serves to indicate that on the left we have $\leq$ and on the right $<$.

The geometric picture corresponding to our situation is



and clearly the interval $[nA, (n+1)A = nA + A)$ has the same length as the segment A (after all, we add A to nA and get $(n+1)A$).

What we have really done is to break up the set of all possible lengths into an infinite collection of disjoint intervals of size A, and any length then falls in exactly one such interval.

The next step is essentially to cut the interval $[nA, (n+1)A)$ in half. More precisely, instead of the two segments $nA < (n+1)A$ we consider the three segments $nA < nA + \frac{1}{2}A < nA + 2(\frac{1}{2}A) = nA + A = (n+1)A$.

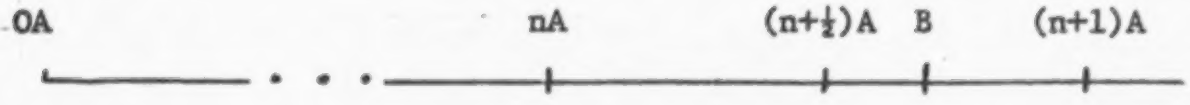
Since $nA \leq B < (n+1)A$ it is clear that exactly one of the possibilities $B < nA + \frac{1}{2}A$ or $B \geq nA + \frac{1}{2}A$ holds; in other words exactly one of the following situations is valid $nA \leq B < nA + \frac{1}{2}A$ or $nA + \frac{1}{2}A \leq B < (n+1)A$. Note that each of the intervals $[nA, nA + \frac{1}{2}A)$, $[nA + \frac{1}{2}A, (n+1)A)$ has size $\frac{1}{2}A$, so that we have improved our knowledge of the length of segment B in the sense that we know in which interval of size $\frac{1}{2}A$ it falls.

In connection with the preceding, we shall also write $nA + \frac{1}{2}A$ as $(n+\frac{1}{2})A$; this is the definition of the symbol $(n+\frac{1}{2})A$ -- until now this symbol had no meaning. If m is a positive integer then it is clear that for any positive integer r the meaning of $\left(\frac{m}{2^r}\right)A$ should be taken as $m(\frac{1}{2^r}A)$. Furthermore, expressions like $\left(\frac{m_1}{2^1} + \frac{m_2}{2^2} + \frac{m_3}{2^3} + \ldots + \frac{m_s}{2^s}\right)A$ should be defined to mean

$\frac{m_1}{2^1}A + \frac{m_2}{2^2}A + \frac{m_3}{2^3}A + \ldots + \frac{m_s}{2^s}A$. In order to keep things consistent it is useful to make some conventions about 0. Thus, we write $nA = nA + oA = (n+o)A = [(n+o(\frac{1}{2})]A = nA + o(\frac{1}{2}A), \ldots, = nA$.

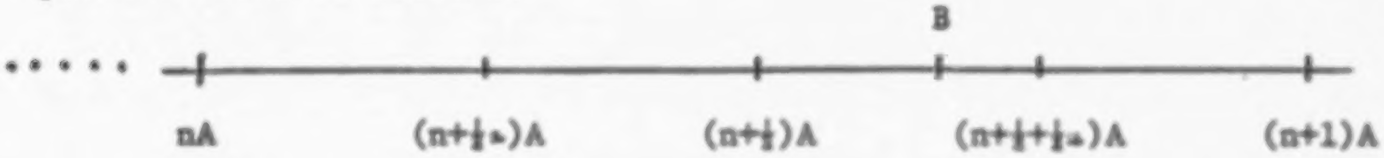
We shall eventually return to a more careful treatment of 0 -- here we merely comment that 0 behaves as expected.

With our new notation in force, we note that the interval $[nA, (n+1)A)$ of size A breaks up into two disjoint intervals $[nA, (n+\frac{1}{2})A)$ and $[(n+\frac{1}{2})A, (n+1)A)$ of size $\frac{1}{2}A$, and that B falls in exactly one of these smaller intervals. Suppose, for purposes of illustration that B falls in the latter interval. The geometric picture then looks as follows:



where the labels of the points signify that they are the endpoints of the segments of that size (all starting from the same point).

Now that we have $(n+\frac{1}{2})A \leq B < (n+1)A$, the same procedure may be repeated. Thus, the interval $[(n+\frac{1}{2})A, (n+1)A)$ of size $\frac{1}{2}A$ breaks up into two disjoint intervals $[(n+\frac{1}{2})A; (n+\frac{1}{2}+\frac{1}{2^2})A)$ and $[(n+\frac{1}{2}+\frac{1}{2^2})A, (n+1)A)$ of size $\frac{1}{2^2}A$, and B falls in exactly one of these intervals. Suppose B falls in the first of these; then the picture looks as follows:



At this stage, in view of assumptions at each step with regard to the location of B, we have

$$nA \leq B < (n+1)A$$

$$(n+\tfrac{1}{2})A \leq B < (n+1)A$$

$$(n+\tfrac{1}{2})A \leq B < (n+\tfrac{1}{2}+\tfrac{1}{2^2})A$$

By our assumption that every segment can be halved, there exists a segment $\frac{1}{2^r}A$ for every positive r and therefore this process of refining our knowledge of the location of B continues indefinitely (in theory). At the first step, B falls in an interval of size A; at the second step, it falls in an interval of size $\frac{1}{2}A$; at the third step, B falls in an interval of size $\frac{1}{2^2}A$; and clearly, at the r^{th} step, B falls in an interval of size $\frac{1}{2^{r-1}}A$. Thus, B falls in each of an infinite sequence of nested intervals (meaning that each interval is contained in the preceding one) -- where the r^{th} interval has size $\frac{1}{2^{r-1}}A$. Since, as seen in the preceding section, the intervals $\frac{1}{2^{r-1}}A$ become as small as desired as r increases, it is clear (intuitively) that there is exactly one "point" that belongs to all the intervals of the nested sequence -- namely, the point which represents the end-point of B. In view of this, it is perfectly natural to say that B is represented by this infinite sequence of nested intervals or inequalities. Conversely, any such infinite sequence of nested inequalities or intervals represents a segment C -- namely, the one whose end-point falls in all the intervals.

The nested sequence of intervals which we have associated with a segment B started with an interval of size A. Since what really matters is that the end-point of B be the unique point which belongs to all the intervals, it does not really matter which interval of the nested sequence is taken as the initial one. In other words, we could throw away the first r intervals of sizes, A, $\frac{1}{2}A$,..$\frac{1}{2^{r-1}}$

and start with the interval of size $\frac{1}{2^r}A$ -- for, after all, this still leaves us with an infinite nested sequence of intervals whose only common point is the end-point of B. It may also be noted that once we have the interval of size $\frac{1}{2^r}A$ then the earlier intervals may be recaptured from it. For example, suppose that at the sixth approximation we know $(17 + \frac{1}{2^2} + \frac{1}{2^4} \frac{1}{2^5})A \leq B < (17 + \frac{1}{2^2} + \frac{1}{2^3})A$ then the intervals preceding this one are:
$(17 + \frac{1}{2^2} + \frac{1}{2^4})A \leq B < (17 + \frac{1}{2^2} + \frac{1}{2^3})A$, $(17 + \frac{1}{2^2})A \leq B < (17 + \frac{1}{2^2} + \frac{1}{2^3})A$, $(17 + \frac{1}{2^2})A \leq B < (17 + \frac{1}{2})A$, $17A \leq B < (17 + \frac{1}{2})A$ $(17)A \leq B < 18A$.

One may ask, at this point, what happens if B turns out eventually to be the same as the left end-point of one of the intervals -- for example, if in the preceding $B = (17 + \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^5})A$? For us, this is nothing more than an accident which does not affect the process; that is, the process still continues and still leads to an infinite nested sequence of intervals which close down on the end-point of B.

Next, let us consider how the relations or operations between segments are reflected in their nested sequences of intervals. Suppose that we have two segments B and C, each expressed in terms of an infinite sequence of nested intervals in terms of A; from these intervals we can decide which is bigger. One simply compares the intervals of corresponding size, and finds the first pair which are not identical -- the one to the right is associated with the bigger segment. For example, suppose that $17A \leq B < 18A$, $(17 + \frac{1}{2})A \leq B < 18A$,....while $17A \leq C < 18A$, $17A \leq C < (17 + \frac{1}{2})A$; clearly, $C < B$. In this type of situation, it is customary to say that we have a lexicographic ordering, because it is essentially like the ordering of words in a dictionary.

What about $B + C$ in terms of the nested intervals? Here one simply takes intervals of corresponding size and adds their end-points. Thus, for the preceding example, we get $34A \leq B + C < 36A$, $(34 + \frac{1}{2})A \leq B + C < (35 + \frac{1}{2})A$,....The nested intervals here are of sizes $2A$, A, $\frac{1}{2}A$, $\frac{1}{2}^{2}A$,....and they do have exactly one point in common -- namely, the end-point of $B + C$.

1-12. Dyadic Expansions

In the preceding section we have seen that once a segment A is fixed then an arbitrary segment B is represented by an infinite sequence of nested intervals (or inequalities) of sizes $\frac{1}{2^n}$ A $n = 0,1,2...$, and conversely. Let us consider a specific example and see what the nested intervals look like. Suppose that we take some segment B and then define the segment A to be A = 3B; this A is to be our fixed segment, and we wish to examine the description of B in terms of A (of course, we are really looking at B = $\frac{1}{3}$ A) as given by nested intervals. Since we wish to work with A and B in concrete fashion they should be taken, at the start, to be neither too big nor too small. When B is compared with A experimentally, we find that the first approximation is

$$0\,A \leq B < 1\,A$$

Now, the procedure for finding the nested intervals associated with B is perfectly straightforward, and if we work carefully and accurately it should turn out that -- the second approximation is

$$0\,A \leq B < \frac{1}{2}\,A$$

while the third approximation is

$$\frac{1}{2^2}\,A \leq B < \frac{1}{2}\,A$$

while the fourth approximation is

$$\frac{1}{2^2}\,A \le B < \left(\frac{1}{2^2} + \frac{1}{2^3}\right) A$$

Continuing, experimentally, in the same way, we find that the fifth approximation is

$$\left(\frac{1}{2^2} + \frac{1}{2^4}\right) A \le B < \left(\frac{1}{2^2} + \frac{1}{2^3}\right) A ,$$

that the sixth approximation is

$$\left(\frac{1}{2^2} + \frac{1}{2^4}\right) A \le B < \left(\frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^5}\right) A$$

and the seventh approximation is

$$\left(\frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6}\right) A \le B < \left(\frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^5}\right) A$$

At this stage, we may well have reached the tolerance limit of our tools; of course, in theory this process of approximation continues ad infinitum.

Our notation is obviously rather cumbersome and it is surely convenient to introduce a more condensed notation. Consider the term $\left(\frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6}\right) A$ from the seventh approximation. If we examine the procedure by which this expression arose, it is clear that its meaning is the same as

$$\left[0 + 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2^2}\right) + 0\left(\frac{1}{2^3}\right) + 1\left(\frac{1}{2^4}\right) + 0\left(\frac{1}{2^5}\right) + 1\left(\frac{1}{2^6}\right)\right] A$$

-- or with the natural use of 0, as

$$\left[0 + \frac{0}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \frac{0}{2^5} + \frac{1}{2^6}\right] A$$

Thus, it is not surprising that we choose to write this as

$$\left[0 \,\middle|\, 0\ 1\ 0\ 1\ 0\ 1\right] A$$

where, as we shall see later, the vertical stroke $\mid$ plays the same role as the decimal point in our usual number system. According to this notation, the seventh approximation looks like

$$\left[0 \,\middle|\, 0\ 1\ 0\ 1\ 0\ 1\right] A \leq B < \left[0 \,\middle|\, 0\ 1\ 0\ 1\ 1\ 0\right] A$$

and, in particular, B falls in the interval of size $\left(\frac{1}{2^6}\right) A$ whose left end-point is $\left[0 \,\middle|\, 0\ 1\ 0\ 1\ 0\ 1\right] A$. Note that from this left end points all the preceding approximations can be recaptured -- the left end-points arise by dropping the right most digits from $\left[0 \,\middle|\, 0\ 1\ 0\ 1\ 0\ 1\right]$ one at a time, and each right hand end-point arises from the corresponding left end-point when we make use of the fact that $\frac{1}{2^r} + \frac{1}{2^r} = \frac{1}{2^{r-1}}$.

More exactly, in our situation we get:

sixth approximation: $\left[0 \,\middle|\, 0\ 1\ 0\ 1\ 0\right] A \leq B < \left[0 \,\middle|\, 0\ 1\ 0\ 1\ 1\right] A$

fifth approximation: $\left[0 \,\middle|\, 0\ 1\ 0\ 1\right] A \leq B < \left[0 \,\middle|\, 0\ 1\ 1\ 0\right] A$

fourth approximation: $\left[0 \,\middle|\, 0\ 1\ 0\right] A \leq B < \left[0 \,\middle|\, 0\ 1\ 1\right] A$

third approximation: $[0\,|\,0\ 1]\ A \le B < [0\,|\,1\ 0]\ A$
second approximation: $[0\,|\,0]\ A \le B < [0\,|\,1]\ A$
first approximation: $0\ A \le B < 1\ A$.

Let us turn, momentarily, from the specific example $B = \frac{1}{3} A$ under consideration, to the general case. Here the segment A is fixed, and B is some fixed, but arbitrary segment. In the first approximation, we have an integer $n \geq 0$ such that $n A \le B < (n+1) A$. The process of subdivision, starting with the interval $[n A, (n+1) A)$, by which we arrive at the infinite sequence of nested intervals associated with B, yields then in the r^{th} approximation an interval of size $\frac{1}{2^{r-1}} A$ in which B lies and whose left hand end-point looks like $\left(n + \frac{a_{-1}}{2} + \frac{a_{-2}}{2^2} + \dots + \frac{a_{-(r-1)}}{2^{r-1}}\right) A = [n\,|\,a_{-1}\ a_{-2}\ \dots\ a_{-(r-1)}] A$ where each of $a_{-1}, a_{-2}, \dots a_{-(r-1)}$ is either 0 or 1. The right hand end-point of this interval is gotten by adding $\frac{1}{2^{r-1}} A$ to the left end-point. Thus, for example, if the left end-point is $[5\,|\,0\ 1\ 0\ 0\ 1\ 1\ 1\ 1]\ A$, then the right end-point is $[5\,|\,0\ 1\ 0\ 1\ 0\ 0\ 0\ 0]\ A$ since we've added /00000001 to 5/01001111. Of course, knowing the left end-point of the r^{th} approximation enables us to determine the left end-point of the $(r-1)^{th}$ approximation -- namely, by simply dropping the a_{r-1} term -- and from it the right end-point of the $(r-1)^{th}$ approximation; so that from the r^{th} approximation, or nested interval, we can recapture all the preceding ones. For example, if as just assumed,

$$[5\,|\,0\;1\;0\;0\;1\;1\;1\;1]\;A \leq B < [5\,|\,0\;1\;0\;1\;0\;0\;0\;0]\;A$$

(this being the 9th approximation) then the 8th approximation is

$$[5\,|\,0\;1\;0\;0\;1\;1\;1]\;A \leq B < [5\,|\,0\;1\;0\;1\;0\;0\;0]\;A\;,$$

and going further we get among others the 4th approximation

$$[5\,|\,0\;1\;0]\;A \leq B < [5\,|\,0\;1\;1]\;A \quad \text{etc....}$$

Returning then to the arbitrary segment B, the r^{th} approximation (that is, the r^{th} nested interval) is determined by its left end-point, which is of form $[n\,|\,a_{-1}\;\; a_{-2}\; \ldots\; a_{-(r-1)}]$ A with each a equal 0 or 1. We therefore have an infinite sequence of these end-points. How are they related? It has already been observed that the $(r-1)^{th}$ end-point is gotten from the r^{th} by dropping the "digit" $a_{-(r-1)}$. In the same way it is clear that if we know the r^{th} end-point, then the next end-point, that is, the $(r+1)^{th}$ (which is the left end-point of the the $(r+1)^{th}$ nested interval) arises by adjoining an extra "digit" (namely, a_{-r}, which is a 0 or a 1) to the representation of the r^{th} end-point.

From all this, we arrive at an infinite sequence of zeros and ones, $a_{-1},\; a_{-2},\; a_{-3},\; a_{-4},$ and may introduce the symbol

2

$$\left[n \mid a_{-1} \;\; a_{-2} \;\; a_{-3} \;\; a_{-4} \cdots \right] A \qquad (*)$$

where the three dots indicate that the expression goes out to infinity. In other words, starting from an arbitrary segment B we are led to associate with it a symbol of type (*). The purpose or meaning of this symbol is simply to provide a simple, compact notation that represents the infinite sequence of nested intervals associated with an arbitrary segment. More precisely, the infinite sequence of nested intervals associated with B determines the symbol (*). Conversely, given an expression of form (*), it determines an infinite sequence of nested intervals -- namely, the r^{th} nested interval has $\left[n \;\; a_{-1} \cdots a_{-(r-1)}\right] A$ as its left end-point, and its size is $\frac{1}{2^{r-1}} A$. Since B is the segment whose right end-point is the unique point which lies in everyone of the infinite sequence of nested intervals associated with B, we are indeed justified in writing

$$B = \left[n \mid a_{-1} \;\; a_{-2} \;\; a_{-3} \cdots \right] A$$

We shall have to learn how to operate with the symbols of form (*). As a matter of fact, at this stage, for given B we do not know how to find the associated symbol (which may be referred to as an infinite dyadic decimal) of form (*). For example, going back to the previous concrete example $B = \frac{1}{3} A$, we have seen that the seventh approximation is

$$[0|0\,1\,0\,1\,0\,1]\,A \leq B < [0|0\,1\,0\,1\,1\,0]\,A$$

Thus, the infinite sequence which gives the dyadic decimal expression for B starts with $0|0\,1\,0\,1\,0\,1$. If our tools are very refined, we may be able to get a few more digits, but it is obvious that we cannot get them all in this way. Based on how things have gone in the first few approximations, one might suspect that for $B = \frac{1}{3} A$

$$B = [0|\underline{0\,1}\ \underline{0\,1}\ \underline{0\,1}\ \underline{0\,1}\ \ldots\ldots]\,A$$

where the notation is designed to indicate that the pair $\underline{0\,1}$ is repeated an infinite number of times -- but this is nothing more than a guess!!

Let us now look at another example. Consider $D = \frac{1}{7} C$ -- that is, choose any segment D and take C = 7 D. If we follow the experimental procedure applied before, then the seventh approximation should turn out to be

$$[0|0\,0\,1\,0\,0\,1]\,C \leq D < [0|0\,0\,1\,0\,1\,0]\,C$$

(There is no need, once this is known, to record the first six approximations.) This leads to a guess that the expansion of D in terms of

C is

$$D = \left[0 \,\middle|\, \underline{0\ 0\ 1}\ \ \underline{0\ 0\ 1}\ \ \underline{0\ 0\ 1}\ \ldots.\right] C$$

We shall return to this question later, and decide if this guess is accurate.

Exercise:

1) Determine the approximations up to and including the seventh order for $B = \frac{1}{5} A$. What is your guess as to the expansion of B with respect to A?

2) Do the same for $C = \frac{6}{5} A$.

3) Do the same for $D = \frac{43}{128} A$.

We conclude this section with one more example. Consider a segment A, and construct a right triangle both of whose legs are segments of length A. Call the hypotenuse B -- we investigate the expansion of B with respect to A. In virtue of the Pythogorean theorem we are really trying to express $B = \sqrt{2}\ A$ in terms of A. Working carefully, we "should" find that the seventh approximation is

$$\left[1 \,\middle|\, 0\ 1\ 1\ 0\ 1\ 0\right] A \leq B < \left[1 \,\middle|\, 0\ 1\ 1\ 0\ 1\ 1\right] A$$

and maybe even that the eighth approximation is

$$\left[1 \middle| 0\,1\,1\,0\,1\,0\,1\right] A \leq B < \left[1 \middle| 0\,1\,1\,0\,1\,1\,0\right]$$

Eventually, this will be seen to provide a very good approximation to the square root of 2.

1 - 13. Dyadic Expansion of Integers

In the preceding section, we have seen that an arbitrary length or segment B can be expressed in terms of a fixed segment A in the form

$$B = \left[n \,\middle/\, b_{-1}\, b_{-2}\, b_{-3} \ldots \right] A$$

where n is an integer greater than or equal to 0, and each b_i for $i = -1, -2, -3, \ldots$ is either 0 or 1. Note that this involves a minute change from the notation used in section 1-12 -- namely, the use of b's instead of a's. It is more logical that, with A fixed and B subject to choice, the expression for B in terms of A should contain b's. In this spirit, for any length C we would write

$$C = \left[m \,\middle/\, c_{-1}\, c_{-2}\, c_{-3} \ldots \right] A$$

with $m \geq 0$ and each c_i equal to 0 or 1 for $i = -1, -2, -3, \ldots$

There is a certain awkwardness and lack of symmetry in the notation for B. On the left side of the vertical stroke we have an integer ≥ 0, and on the right side an infinite sequence of zeros and ones. Can something be done to make the left side also consist only of zeros and ones so that both sides of the vertical strokes are similar objects and can then be treated in unified fashion? Thus, we really wish to examine the case

$$B = nA = \left[n \,\middle/\, 0\; 0 \ldots 0 \ldots \right] A$$

-- that is, where the sequence of nested intervals expressing B in terms of A all have n A, with n an integer ≥ 0, as the left-hand end-point. We would hope to be able to replace n by a sequence of 0's and 1's. In an expression of form $\left[n \,\middle|\, b_{-1}\ b_{-2}\ b_{-3} \ldots \right]$ the meaning of the stuff to the right of the vertical stroke is, of course,

$$\frac{b_{-1}}{2} + \frac{b_{-2}}{2^2} + \frac{b_{-3}}{2^3} + \ldots + \frac{b_{-r}}{2^r} + \ldots$$

or, what is the same

$$b_{-1}2^{-1} + b_{-2}2^{-2} + b_{-3}2^{-3} + \ldots + b_{-r}2^{-r} + \ldots$$

Thus, we may say somewhat carelessly that the stuff to the right of the vertical stroke represents a "sum of powers of 2" -- namely, negative ones. We shall try to express n, the stuff to the left of the vertical stroke, as a sum of powers of 2; if this can be done, we would expect to use only non-negative powers of 2.

Let us start with some simple concrete examples. Consider n = 27 and $B = 27A = \left[27 \,\middle|\, 0\ 0\ \ldots\ 0\ \ldots \right] A$. It is not hard to see that

(*)

$$\begin{aligned} 27 &= 2^4 + 2^3 + 2 + 1 \\ &= 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 \end{aligned}$$

Thus, we may associate with 27 the "sequence" 11011 determined by (*), and write

$$B = 27A = \left[11011 / 00\ldots0\ldots \right] A$$

Consider next n = 69 and B = 69A; then it may be observed that

$$69 = 2^6 + 2^2 + 1$$

$$= 1\cdot2^6 + 0\cdot2^5 + 0\cdot2^4 + 0\cdot2^3 + 1\cdot2^2 + 0\cdot2^1 + 1\cdot2^0$$

Thus, 1000101 is a sequence of zeros and ones to be associated with 69, and we write

$$B = 69A = \left[1000101 / 0\ldots0\ldots \right] A$$

Finally, let us consider n = 84 and B = 84A. Since

$$84 = 2^6 + 2^4 + 2^2$$

$$= 1\cdot2^6 + 0\cdot2^5 + 1\cdot2^4 + 0\cdot2^3 + 1\cdot2^2 + 0\cdot2^1 + 0\cdot2^0$$

we may write

$$B = 84A = \left[1010100 / 0\ldots0\ldots \right] A$$

From these examples we may conjecture that if n is a positive integer and B = nA, then we can write

$$(\#) \qquad n = b_r \cdot 2^r + b_{r-1} \cdot 2^{r-1} + \ldots + b_1 \cdot 2^1 + b_0 \cdot 2^0$$

where each $b_r, b_{r-1}, \ldots, b_1, b_0$ is 0 or 1, $b_r = 1$, and $r \geq 0$, and express B by the notation

$$B = nA = \left[b_r\ b_{r-1}\ \ldots b_1\ \ b_0 \middle/ 0\ldots0\ldots \right] A$$

The crucial question, therefore, is whether or not any positive integer n can be expressed in the form (#) -- that is, as a sum of powers of 2 -- more precisely, can n be written as a sum of certain of the integers $2^0 = 1$, $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, $2^6 = 64$, $2^7 = 128, \ldots$

Exercise: Can you express each of the following integers as a sum of powers of 2? -- 78, 99, 129, 150, 250, 437, 500.

By now, the reader is probably convinced that every positive integer can indeed be written as a sum of powers of 2. Let us try to indicate informally why this appears to be true. Using only $2^0=1$ and $2^1=2$, we can express 1, 2, 3 (but not 4) in the appropriate form. Thus, if we throw in $2^2=4$, then using 2^0, 2^1, 2^2, we can express (in addition to 1, 2, 3) 4, 5, 6 and 7 as a sum of powers of 2. Throwing in $2^3=8$, we can then express 1 thru 7 and also 8 thru $8 + 7=15$ in the desired form. Thus proceeding inductively -- given $2^0, 2^1, \ldots, 2^r$ we can express every integer from 1 to $2^{r+1} - 1$ as a sum of certain of the preceeding powers of two.

Problem:

Show that $2^0 + 2^1 + 2^2 + 2^3 + \cdots + 2^r = 2^{r+1} - 1$

For a complete discussion of writing a decimal integer as powers of 2 the reader is directed to Appendix I.

One may now inquire if the base 2 expansion of any integer n is unique for example the decimal number 179. Of course, the method of dividing by 2 and using the remainders leads to a single result, but this does not, in itself guarantee that there cannot exist some expression for 179 other than 10110011. To show that the expansion of an integer n is unique we suppose that there are two such expansions and show that they must be identical. Thus suppose that

$$n = b_r b_{r-1} \dots b_0 \qquad\qquad b_r = 1$$

and also that

$$n = c_s c_{s-1} \dots c_0 \qquad\qquad c_s = 1$$

we must show that $r = s$ and that for $i = 0, 1, \dots, r$, $c_i = b_i$.

The hypotheses say that

$$(\#) \qquad b_r 2^r + b_{r-1} 2^{r-1} + \dots + b_1 \cdot 2 + b_0 = c_s 2^s + c_{s-1} 2^{s-1} + \dots + c_1 \cdot 2 + c_0$$

Can it be that $b_0 \neq c_0$? If so, then say $c_0 > b_0$, which means that $c_0 = 1$ and $b_0 = 0$. But this says that n, as given by the left side of (#), is even (because 2 divides the left side) while n, as given by the right side of (#), is odd (because division by 2 gives a remainder of 1) -- a contradiction. We conclude that we must have $b_0 = c_0$. Consequently, upon subtracting or removing $b_0 = c_0$ from both sides of (#) and then dividing the result by 2, we arrive at

$$b_r 2^{r-1} + b_{r-1} 2^{r-2} + \dots + b_2 \cdot 2 + b_1 = c_s 2^{s-1} + \dots + c_2 \cdot 2 + c_1$$

But this is exactly the same set-up as (#), and as was done there we conclude that $b_1 = c_1$. This process may be repeated inductively, to get $b_i = c_i$. If the b's are used up first so that $b_0 = c_0$, $b_1 = c_1$, ..., $b_r = c_r$ then it follows easily that $c_{r+1} = \dots = c_s = 0$,

and indeed $r = s$.

The upshot of this entire discussion is that instead of expressing an arbitrary segment B in terms of the fixed segment A in the form $B = \left[n \,/\, b_{-1}\ b_{-2} \dots \right] A$, we may write

$$B = \left[b_r\ b_{r-1} \dots b_0 \,/\, b_{-1}\ b_{-2} \dots \right] A$$

where $b_r = 1$ and each b_i is 0 or 1. How are all the b's to be found? As of now we first locate B in an interval of size A, $B \in [nA, (n + 1) A)$, -- that is,

$$nA \leq B < (n + 1) A$$

This determines n, and then expressing n in base 2 gives $b_r, \ldots, b_0$. Furthermore, by repeated halving of the interval $[nA, (n + 1) A)$ there arises a nested sequence of intervals described completely by $b_{-1}, b_{-2}, b_{-3}, \ldots$. The approach to the left side of the vertical stroke differs from the approach to the right side, but it is important to observe that it need not be so -- we can treat all the b's, rather than just those on the right of the vertical stroke, according to the same nested sequence of intervals procedure that was used before. More precisely, if $B \in [0A, 1A)$ then dividing intervals in half in the usual way we get $B = [0 | b_{-1} \; b_{-2} \ldots] A$. In the general case where $B \geq A$, consider the lengths $A = 2^0 A$, $2A$, $2^2 A = 4A$, $2^3 A = 8A$, ..., $2^m A$, ... -- these lengths get arbitrarily large, and there exists a unique integer $r \geq 0$ such that

$$B \in [2^r A, 2^{r+1} A)$$

The size of this interval in which B falls is $2^r A$, and we have

$$2^r A \leq B < 2^{r+1} A$$

For this r, we put $b_r = 1$. Then halving this interval B falls in exactly one of the intervals of size $2^{r-1} A$

$$\left[2^r A, (2^r + 2^{r-1}) A\right), \quad \left[(2^r + 2^{r-1}) A, 2^{r+1} A\right)$$

If B falls in the first one, we have $b_{r-1} = 0$, if it falls in the second one, then $b_{r-1} = 1$. In any case, the canonical method for deriving a nested sequence of intervals applies and gives us, starting from $b_1 = 1$, all the b's so that

$$B = \left[b_r \, b_{r-1} \ldots b_0 \,\middle|\, b_{-1} \, b_{-2} \ldots\right] A$$

Of course, $b_r \, b_{r-1} \ldots b_0$ still represents the integer n such that $nA \leq B < (n + 1) A$.

1-14: Computation with Dyadic Expansions

We know that if a segment A is fixed then any segment B can be expressed in terms of A in the form B = βA where

$$\beta = \left[b_r \; b_{r-1} \; \ldots \; b_0 \,/\, b_{-1} \; b_{-2} \; \ldots \right]$$

-- in other words, β may be considered as an infinite sequence of zeros and ones with one "spot" distinguished from all others, namely, by the vertical stroke. In such a situation one may say (with some degree of carelessness) that β is the dyadic expansion of B and that β is a "real number". It is time to learn how to operate and compute with such dyadic expansions, and of course the derivation of computation rules must arise from the rules for operating with segments.

Thus, suppose we have another segment C; then C may be expressed in the form C = γA, where

$$\gamma = \left[c_s \; \ldots \; c_0 \,/\, c_{-1} \; c_{-2} \; \ldots \right] \qquad c_i = 0 \text{ or } 1$$

Now, the sum of the two segments B + C = C + B also has an expression in terms of A -- say B + C = δ A where

$$\delta = \left[d_t \cdots d_0 / d_{-1}\ d_{-2} \cdots \right]$$

Our first objective then is to add β and γ,

$$\left[b_r \cdots b_0 / b_{-1}\ b_{-2} \cdots \right] + \left[c_s \cdots c_0 / c_{-1}\ c_{-2} \cdots \right]$$

and in this way to find δ For this, it is convenient and instructive to start working with integers -- explicitly, suppose that $\beta = m$, $\gamma = n$ are integers, $B = mA$, $C = nA$, $\beta = m = \left[b_r \cdots b_0 / 00 \ldots 0 \ldots \right]$, $\gamma = n = \left[c_s \ldots c_0 / 0 \ldots 0 \ldots \right]$, so that $B + C = (m + n)\ A$ and $\delta = m + n = \left[d_t \cdots d_0 / 0 \ldots 0 \ldots \right]$. In view of the fact that

$$\left[b_r \ldots b_0 / 0 \ldots \right] \text{ represents } b_r 2^r + b_{r-1} 2^{r-1} + \ldots + b_1 \cdot 2 + b_0$$

and

$$\left[c_s \ldots c_0 / 0 \ldots \right] \text{ represents } c_s 2^s + c_{s-1} 2^{s-1} + \ldots + c_1 \cdot 2 + c_0$$

with each $b_r, \ldots, b_0, c_s, \ldots, c_0$ being 0 or 1, in order to carry out the addition

$$\left[b_r \dots b_0 / 0 \dots \right] + \left[c_s \dots c_0 / 0 \dots \right]$$

we need only carry out the addition of their base 2 representations -- and this is trivial because we know how to add ordinary integers. Rather than letting ourselves get bogged down in all the verbiage needed to describe accurately how one adds two integers in their base 2 representations in general, it is better to turn to a few examples which will serve to illustrate the points involved.

Consider $\beta = m = 27$ and $\gamma = n = 69$. We recall that $27 = \left[11011 / 0 \dots \right]$ which reflects the fact that

$$27 = 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1$$

and write all this simply as 11011. In the same way, $69 = \left[1000101 / 0 \dots \right] = 1000101$, which expresses symbolically the relation

$$69 = 1 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2 + 1$$

In order to find the sum 11011 + 1000101 we take

$$(1\cdot 2^4 + 1\cdot 2^3 + 0\cdot 2^2 + 1\cdot 2^1 + 1) + (1\cdot 2^6 + 0\cdot 2^5 + 0\cdot 2^4 + 0\cdot 2^3 + 1\cdot 2^2 + 0\cdot 2^1 + 1)$$

and according to the usual rules for adding integers this equals

$$(*) \qquad 1\cdot 2^6 + 0\cdot 2^5 + 1\cdot 2^4 + 1\cdot 2^3 \; 1\cdot 2^2 + 1\cdot 2^1 + 2\cdot 2^0$$

Unfortunately, this expression is not quite in the proper form of a base 2 expansion -- we must have a sequence of zeros and ones, but the last coefficient (the one associated with the 2^0 term) here is 2. Of course, it is not hard to adjust (*) to put it in the correct form. First of all, $2\cdot 2^0 = 2\cdot 1 = 1\cdot 2 = 1\cdot 2^1$, so that the $2\cdot 2^0$ term may be replaced by $1\cdot 2^2$, thus giving $1\cdot 2^1 + 1\cdot 2^1 = 2\cdot 2^1$ -- and (*) may then be re-written as

$$1\cdot 2^6 + 0\cdot 2^5 + 1\cdot 2^4 + 1\cdot 2^3 + 1\cdot 2^2 + 2\cdot 2^1 + 0\cdot 2^0$$

But now we have the analogous difficulty with $2\cdot 2^1$; however, this term may be replaced by $1\cdot 2^2$ -- to give

$$1\cdot 2^6 + 0\cdot 2^5 + 1\cdot 2^4 + 1\cdot 2^3 + 2\cdot 2^2 + 0\cdot 2^1 + 0\cdot 2^0$$

The method is now clear, if we ever get $2\cdot 2^r$ this may be replaced by $1\cdot 2^{r+1}$, so that things are "moved" one step to the left. Of course, if a term $3\cdot 2^{r}$ should arise it needs to be replaced by $1\cdot 2^{r+1} + 1\cdot 2^{r}$. Continuing in this way until all terms have coefficient either 0 or 1, it is easy to see that the end result is

$$1\cdot 2^{6} + 1\cdot 2^{5} + 0\cdot 2^{4} + 0\cdot 2^{3} + 0\cdot 2^{2} + 0\cdot 2^{1} + 0\cdot 2^{0}$$

This says that

$$11011 + 1000101 = 1100000$$

which, it may be noted, is another way of saying that

$$27 + 69 = 96.$$

Exercise: 1) Verify that 11011 + 1010100 = 1101111, and that this corresponds to 27 + 84 = 111 (one hundred and eleven).

2) Verify that 1000101 + 1010100 = 10011001 and that this corresponds to 69 + 84 = 153.

3) Find the following sums and the ordinary integers to which they correspond

11011 + 10110011 =

1000101 + 10110011 =

1010100 + 10110011 =

10110011 + 10110011 =

4) Perform the following additions and check the results by translating everything to ordinary integers:

$$10110 + 110001 =$$

$$11001101 + 100110 =$$

$$111111101 + 101101 =$$

5) Perform the following additions by changing to base 2, adding and then re-writing the result as an ordinary integer:

$$98 + 47 =$$

$$198 + 943 =$$

$$7511 + 5751 =$$

For those who are experiencing difficulty with the above problems may we recommend a review of Appendix I and serious study of Appendix II.

We still want to be able to add in the general case, namely, where each of our dyadic expansions

$$\beta = \left[b_r \ldots b_0 / b_{-1}\ b_{-2} \ldots \right] \text{ and } \gamma = \left[c_s \ldots c_0 / c_{-1}\ c_{-2} \ldots \right]$$

is really infinite (this means that an infinite number of the b's are 1 and also that an infinite number of the c's are 1). However, this will be deferred until later. On the other hand, it is clear that if both dyadic expansions are finite, then exactly the same principles used in adding dyadic expansions of integers apply. The only variation is the use of the vertical stroke(in both directions) for indexing purposes. Thus, for example, we can add

$$\left[101101 / 1010100 \ldots \right] + \left[11001 / 0011000 \ldots \right]$$

without any difficulty via

$$
\begin{array}{cccccccc|ccccc}
 & & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
+ & & & & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
\hline
 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1
\end{array}
$$

-- so the result is $\left[1000110 / 110110 \ldots \right]$, and we do not find it necessary to re-interpret the symbols involved as numbers, fractions or powers of 2 in order to perform the mechanical act of computation.

Exercise: a) Perform the indicated additions and translate the results to ordinary rational numbers (for example, in the preceding,

$$\left[101101 / 10101 \right] + \left[11001 / 0011 \right] = \left[1000110 / 11011 \right]$$

translates to 45 21/32 + 25 3/16 = 70 27/32)

$$\left[1101\ 0101 \right] + \left[100110 / 11101 \right]$$

$$\left[111101 / 11001 \right] + \left[1011001 / 1011101 \right]$$

b) Perform the following additions of rational numbers by finding their dyadic expansions, adding these, and then translating the result to rational form:

$$79\ 19/32 \quad + \quad 85\ 31/64$$

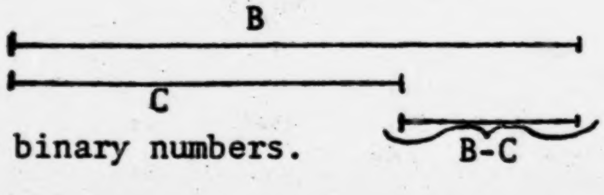
$$157\ 17/32 \quad + \quad 193\ 111/128$$

~~Having dealt with the addition of dyadic expansions (at~~ least, in the situation where both expansions are finite) we turn to the question of subtraction. More precisely, given segments B, C expressed in terms of A as B = β A, C = γ A then

$$B - C = (\beta - \gamma)\ A$$

so that for $\beta = \left[b_r \ldots b_0 / b_{-1} \ldots \right]$, $\gamma = \left[c_s \ldots c_0 / c_{-1} \ldots \right]$ we need to learn how to compute $\beta - \gamma$, the dyadic expansion of B - C in terms of A.* Of course, before this can be discussed it is essential that C $<$ B, so as a first step we should learn to recognize which of two dyadic expansions β and γ is the bigger one. In order to do this, it is convenient to start with the simple situation where both β and γ are integers --

* Point of clarification -- B - C represents the segment we would "attach" to C so that their combined lengths [c + (B-C)] would be as long as B



Whereas B and γ are binary numbers.

$$\beta = m = \left[b_r \dots b_0 \,/\, 0 \dots \right] \qquad \gamma = n = \left[c_s \dots c_0 \,/\, 0 \dots \right]$$

(as a matter of fact, this special case includes the necessary general principles). As some examples, we note that

$$10 > 1, \quad 100 > 11, \quad 1000 > 111, \quad 10000 > 1111, \dots \text{etc.}$$

-- in fact, according to the way addition works, adding 1 to the right side of each of these gives the left side. Therefore, the inequalities are as stated. Even more, the left side is in each case the expression for 2^r $(r \geq 1)$, so it follows immediately that the right side is $2^r - 1$ whose expression is

$$2^r - 1 = 1 \cdot 2^{r-1} + 1 \cdot 2^{r-2} + \dots + 1 \cdot 2 + 1$$

In other words, $111\dots1 = b_r\, b_{r-1} \dots b_0$ is the expansion of $2^{r+1} - 1$.

Now, it is clear that 11111...1 (with t 1's) is greater than any finite expansion with the same number of terms (provided, of course, that at least one of the terms is 0), and by transitivity we see that 1000...0 (with t 0's) is greater than any expansion with t terms. It is equally clear that any expression with t + 1 terms starting with a 1 is greater than any expression with t terms.

This constitutes a large quantity of words for a general principle which can best be understood from a few examples:

$$100 > 10, \quad 10101 > 1101, \quad 10001 > 111, \quad 11111 > 1110, \quad 111000 > 11011$$

According to this, if we have two integers expressed dyadically $\beta = [b_r \, b_{r-1} \ldots b_0 / 0 \ldots]$ and $\gamma = [c_s \, c_{s-1} \ldots c_0 / 0 \ldots]$ with both leading terms b_r and c_s equal to 1 then

$$r > s \Longrightarrow \beta > \gamma$$

-- in short, if one of two binary expressions for integers (with both expressions starting with 1) is longer than the other, then the corresponding integer is the larger of the two.

It remains to examine the situation where $r = s$ -- that is, where the two expressions have the same length. In this case, $b_r = c_r = 1$, and the leading terms of β and γ may be ignored -- in other words, it suffices to compare $b_{r-1} \ldots b_0$ and $c_{r-1} \ldots c_0$, which are of the same length except that the leading terms can be 0. Thus, if $b_{r-1} \neq c_{r-1}$ the matter is settled, while if

$b_{r-1} = c_{r-1}$ we continue by discarding these terms. In summary, for expansions of equal length, we work from left to right, find the first place where the two expressions differ (that is, where they have different digits) and the one which has a 1 in this place is bigger than the one that has a 0. Some examples of this are:

$11010 > 10001 \qquad 11011 > 11001 \qquad 110011010110 > 1110011001110$

__Exercise__: Line up the following dyadic expansions of integers according to size:

101, 1011, 11110001, 1010110, 110, 1101, 1001,

100100, 1101011, 101101, 10110111, 101100.

We can now decide which of two integers in dyadic form is greater (note that we read this off directly from the notation and do not have to evaluate the symbol as an ordinary integer), and because of our knowledge of addition the procedure for doing subtraction is straightforward. Thus, if $\beta = 10101$ and $\gamma = 111101$ then surely $\gamma > \beta$ and $\gamma - \beta = 101000$. The only surprise or difficulty is perhaps that in "borrowing" one may end up doing the real borrowing from far up the line -- for example,

$$\begin{array}{r} 1\,0\,1\,1\,0\,0\,1\,0 \\ -\qquad 1\,1\,1\,0\,1\,1\,1 \\ \hline 1\,1\,1\,0\,1\,1 \end{array}$$

The reader surely understands how to borrow and subtract, and long-winded explanations at this point would most likely be confusing. We shall merely indicate one way in which one might do subtraction without the mental effort of keeping track of the borrowings. In the example, and whenever subtraction is to be performed, we would like to re-write the top number, permitting digits other than 0, 1, so that the term in each column is the term in the corresponding column of the bottom number. This involves, when appropriate, replacing 10 by 02, 100 by 12, 1000 by 112, 10000 by 1112 etc. Thus, in the example, 10110010 = 10000000 + 110010, and the first of the terms on the right equals 1111111 + 1 which we have permitted ourselves the luxury of writing as 1111112. Consequently, since 1111112 + 110010 equals 1221122, and the subtraction

$$\begin{array}{r} 1\,0\,1\,1\,0\,0\,1\,0 \\ -\quad 1\,1\,1\,0\,1\,1\,1 \\ \hline \end{array}$$

becomes

$$\begin{array}{r} 1\,2\,2\,1\,1\,2\,2 \\ -\quad 1\,1\,1\,0\,1\,1\,1 \\ \hline \end{array}$$

and the result is obviously 111011.

Exercise: Perform the following subtractions and interpret them in terms of ordinary integers:

110 - 101, 10010 - 1111, 1010101 - 1010, 11100100010 - 1101101110.

Once we know how to compare the size of two integers expressed in dyadic form and how to subtract one from the other, it is easy to see that some procedures apply for any finite dyadic expansions. Thus, to compare

$$\beta = \left[b_r \dots b_0 \middle| b_{-1} \dots b_{-p}\ 0 \dots \right] \text{ and } \gamma = \left[c_s \dots c_0 \middle| c_{-1} \dots c_{-q}\ 0 \dots \right]$$

the vertical stroke plays a key role. Clearly, if $b_r \ldots b_0 >$ $c_s \ldots c_0$ (these are expansions of integers, and according to the preceding we can decide which is bigger) then $\beta > \gamma$. On the other hand, if $b_r \ldots b_0 = c_s \ldots c_0$ (that is, if the left sides of the vertical stroke are identical) then working to the right of the vertical stroke, from left to right, we locate the first place where $b_t \neq c_t$. One of these is 1 (and the other is 0), and this determines the bigger of the two dyadic expansions.

As for subtracting, this goes exactly as for integers; for integers, the vertical stroke is at the right edge for both terms, while for finite dyadic expressions the vertical strokes must be lined so, and from this starting place all the columns are then lined up.

Exercise: Line up all of the following finite dyadic expansions according to size, and then subtract each from the next largest one.

[11/011], [1001/10], [11011/0101], [101/101], [1101/01101],

[10/101], [110110/01], [111/00001], [101/00101],

[10001/10001], [110/011], [1010/10101]

Our discussion throughout this section has been rather formal and algebraic (rather than geometric) and in this spirit let us practice multiplication in simple cases even though a geometric interpretation is given in the laboratory exercises. Consider two integers $m = \left[b_r \ldots b_0 / 0 \ldots \right]$, $n = \left[c_s \ldots c_0 / 0 \ldots \right]$ and their dyadic expansions -- then, of course, $mn = (b_r 2^r + \ldots + b_1 \cdot 2 + b_0)$ $(c_s 2^s + \ldots + c_1 \cdot 2 + c_0)$ and by using the well-known rules for computation with integers we can get the right side in form

$$d_{r+s} 2^{r+s} + \ldots + d_1 \cdot 2 + d_0$$

so that $mn = \left[d_{r+s} \ldots d_0 / 0 \ldots \right]$. Let us show how to carry out these steps, in practice, using our compact notation by considering an example. Suppose $m = \left[101101 / 0 \ldots \right]$, $n = \left[11011 / 0 \ldots \right]$, since $11011 = 10000 + 1000 + 10 + 1$ we have at the start

$$(101101) \times (11011) = (101101) \times (10000 + 1000 + 10 + 1)$$

$$= (101101) \times (10000) + (101101) \times (1000) + (101101) \times (10)$$

$$+ (101101) \times (1)$$

Therefore, everything boils down to multiplication by numbers of form 1, 10, 100, 1000, 10000, 100000, ... and then performing some additions (and this observation obviously applies whenever we wish to multiply any two integers). Now multiplication by 1 is trivial; the key is multiplication by 10. Since 10 represents the integer 2, multiplication of a number by 10 means multiplying it by 2, which in turn means adding the number to itself. Thus,

$$(101101) \times (10) = 101101 + 101101 = 1011010$$

We see, from what happens when a dyadic expansion of an integer is added to itself, that here and in general, multiplication by 10 involves moving everything over one place to the left -- or more precisely placing a 0 at the end of the dyadic expansion.

Turning to multiplication by 100, we note that 100 = (10) x (10) (which we may also write as $(10)^2$) so that multiplication by 100 involves multiplying by 10 twice -- it is therefore accomplished by placing two 0's at the end of the dyadic expansion of the number we are multiplying. Repeating the process, it follows that multiplication by 100 ... 0, with r zeros (so this is (10) x (10) ... x (10) r times -- i. e. $(10)^r$) means multiplying by (10) r times, and involves placing r zeros at the end.

Coming back to our example we see that

$$(101101) \times (11011) = 1011010000 + 101101000 + 1011010 + 101101$$

and all the work can be coalesced and done directly via the familiar type of line-up:

$$\begin{array}{rrl}
 & 1\,0\,1\,1\,0\,1 & \\
\times & \underline{1\,1\,0\,1\,1} & \\
\hline
 & 1\,0\,1\,1\,0\,1 & \text{(mult. by 1)}\\
 & 1\,0\,1\,1\,0\,1\,0 & \text{(mult. by 10)}\\
 & 1\,0\,1\,1\,0\,1\,0\,0\,0 & \text{(mult. by 1000)}\\
 & 1\,0\,1\,1\,0\,1\,0\,0\,0\,0 & \text{(mult. by 10000)}\\
\hline
 & 1\,0\,0\,1\,0\,1\,1\,1\,1\,1\,1 &
\end{array}$$

Note that we do not even have to do the multiplications from right to left; any order will do provided the columns are lined up correctly.

Exercise: a) Perform the following multiplications, and check them by translating to "ordinary integers".

(10101) x (110); (1011101) x (11001); (11011011) x (101101111)

b) Verify the distributive law in the following examples, by performing all the operations with dyadic expansions:

179 (47 + 137) = (179)(47) + (179)(137)

(5311)(2132 + 1897) = (5311)(2132) + (5311)(1897).

Finally, to conclude this section, it remains to examine the multiplication of two finite dyadic expansions

$$\beta = \left[b_r \ldots b_0 / b_{-1} \ldots b_{-p} \right], \quad \gamma = \left[c_s \ldots c_0 / c_{-1} \ldots c_{-q} \right]$$

There are really no new principles involved, one merely keeps track of the vertical stroke. In particular, here, multiplication by 10 = $[10/0 \ldots]$ requires moving the vertical stroke one place to the right, and multiplication by 100 = $[100/0 \ldots]$ requires moving the vertical stroke two places to the right, ...etc... (Note that this is exactly what was involved when multiplying integers -- for example, (1011) x 100 = $[1011/00 \ldots]$ = $[101100/0 \ldots]$ = 101100).

What about multiplication by [0/10 ...], [0/010 ...], [0/0010 ...], and so on? We know that [0/10 ...] is the representation of ½ (and we may also write it as $(10)^{-1}$, and it is *obvious* that multiplying by ½ requires moving the vertical stroke one place to the left (after all, when the result of multiplying by ½ is then multiplied by 2 = 10 we are back where we started). Furthermore, to multiply by [0/01] = $1/2^2$ = [0/10] x [0/10] we must clearly move the vertical stroke two places to the left, ... and so on.

As an illustration, consider:

[1 0 1 1 1 0/0 1 1 0 1] x [1 0 0 1/1 0 1 1] —

$$\begin{array}{rrl}
 & 1\,0\,1\,1\,1\,0/0\,1\,1\,0\,1 & \\
\times & \underline{1\,0\,0\,1|1\,0\,1\,1} & \\
 & 1\,0|1\,1\,1\,0\,0\,1\,1\,0\,1 & (\text{mult. by } 10^{-4}) \\
 & 1\,0\,1|1\,1\,0\,0\,1\,1\,0\,1 & (\text{mult. by } 10^{-3}) \\
 & 1\,0\,1\,1\,1|0\,0\,1\,1\,0\,1 & (\text{mult. by } 10^{-1}) \\
 & 1\,0\,1\,1\,1\,0|0\,1\,1\,0\,1 & (\text{mult. by } 1 = 10^{0}) \\
 & 1\,0\,1\,1\,1\,0\,0\,1\,1|0\,1 & (\text{mult. by } 10^{3}) \\
\hline
 & [1\,1\,1\,0\,0\,0\,0\,0\,1|1\,0\,0\,0\,1\,1\,1\,1\,1] &
\end{array}$$

-- we may leave it to the reader to check that the above is one way of showing that

(46 13/32) x (9 11/16) = 449 287/512

Exercise: a) Compute [11010/11011] x [1011/1011] and check by transferring to "ordinary" notation.

b) Compute (79 22/32) x (58 15/32), and check by transferring to dyadic notation.

1-15. Computation with Infinite Dyadic Expansions

If A is a fixed segment then we recall that every segment B can be expressed in the form $B = \beta A$ where $\beta = [b_r \ldots b_0 | b_{-1} \ldots]$ is an infinite dyadic expression - in other words, in general, an infinite number of the digits b_i are 1. It is with such objects β that we wish to compute; in the preceding section, we learned how to compute with certain special kinds of β's, namely, the finite ones - that is, those with only a finite number of digits b_i equal to 1, and which could therefore be expressed using only a finite number of b's. Of course, even though β has an infinite expansion this does not mean that we know all the b_i's, or that we have a rule which enables us to find every b_i. (This is entirely analogous to the fact that the number $\pi = 3.14159\ldots$ is an infinite decimal which, with the advent of computers, we now know up to some 2000 places; however, the remaining infinite number of digits are not known.) Thus, it is not surprising that to compute numerically with an infinite expression β, (all of whose terms may not even be known to us) we work with finite approximations to β, and that the geometric aspects play an important role.

Consider two segments $B = \beta A$, $C = \gamma A$ where $\beta = [b_r \ldots b_0 | b_{-1} \ldots]$ and $\gamma = [c_s \ldots c_0 | c_{-1} \ldots]$. It will be convenient to write

$\beta_0 = [b_r \ldots b_0 | 000 \ldots]$

$\beta_1 = [b_r \ldots b_0 | b_{-1} 000 \ldots] = \beta_0 + b_{-1} \cdot \frac{1}{2}$

$\beta_2 = [b_r \ldots b_0 | b_{-1} b_{-2} 000 \ldots] = \beta_1 + b_{-2} \cdot \frac{1}{2^2} = \beta_0 + b_{-1} \cdot \frac{1}{2} + b_{-2} \cdot \frac{1}{2^2}$

.

.

.

$\beta_t = [b_r \ldots b_0 | b_{-1} b_{-2} \ldots b_{-t} 000 \ldots]$ for $t = 0, 1, 2, 3, \ldots$

Thus, β_t is the finite dyadic approximation to β gotten by using all the digits up to and including the $t^{\underline{th}}$ place to the right of the vertical stroke, and of course

$$\beta_t = \beta_{t-1} + b_{-t} \cdot \frac{1}{2^t}$$

Naturally, the same notation applies to γ, and we have

$$\gamma_t = [c_s \ldots . c_0 | c_{-1}\ c_{-2} \ldots c_{-t}\ 00 \ldots] \qquad t = 0,1,2,3,\ldots .$$

We recall further that βA (and γA too) is really a short-hand notation for an infinite sequence of nested intervals; in fact, in virtue of the discussion in sections 1-12 and 1-13 combined with the notation here, we have $B = \beta A$ and $C = \gamma A$ given by the nested sequences:

$$\beta_0 A \leq B < (\beta_0+1)A \qquad \gamma_0 A \leq C < (\gamma_0+1)A$$

$$\beta_1 A \leq B < (\beta_1+\tfrac{1}{2})A \qquad \gamma_1 A \leq C < (\gamma_1+\tfrac{1}{2})A$$

$$\beta_2 A \leq B < (\beta_2+\tfrac{1}{2^2})A \qquad \gamma_2 A \leq C < (\gamma_2+\tfrac{1}{2^2})A$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\beta_t A \leq B < (\beta_t+\tfrac{1}{2^t})A \qquad \gamma_t A \leq C < (\gamma_t+\tfrac{1}{2^t})A$$

$$\beta_{t+1} A \leq B < (\beta_{t+1}+\tfrac{1}{2^{t+1}})A \qquad \gamma_{t+1} A \leq C < (\gamma_{t+1}+\tfrac{1}{2^{t+1}})A$$

$$\vdots \qquad\qquad\qquad \vdots$$

These infinite sequences of nested intervals provide complete descriptions of B and C respectively, and they provide us with the theoretical tools for discussing computations. The easiest case — when the sequence of nested intervals is finite, which means that $\beta = [b_r \ldots b_0 | b_{-1} \ldots b_{-p}\, 00\ldots] = \beta_p$ — occurs when B falls on the left hand end-point of one of the nested intervals; this case was treated in the preceding section, and the infinite case is handled by making use of finite cases which approximate it.

First of all, we may note in passing, how the infinite dyadic expansions may be used to compare the size of B and C. If $\beta_0 > \gamma_0$ (these are finite dyadic expressions, and we already know how to compare them) this means that $\beta_0 A \leq B < (\beta_0+1)A$, $\gamma_0 A \leq C < (\gamma_0+1)A$, so that B falls in the interval $[\beta_0, \beta_0+1)A$ which is entirely to the right of the interval $[\gamma_0, \gamma_0+1)A$ in which C falls — consequently $B > C$, and we also write $\beta > \gamma$. On the other hand, if $\beta_0 = \gamma_0$, then we compare β_1 and γ_1; if $\beta_1 > \gamma_1$, then as above $B > C$; and if $\beta_1 = \gamma_1$ then the process is repeated with β_2 and γ_2. We proceed therefore until we arrive at some subscript p where $\beta_p \neq \gamma_p$, and this settles the decision for us. Of course, it is clear from the geometry of nested intervals, that if $\beta_i > \gamma_i$ then $\beta_{i+1} > \gamma_{i+1}$, $\beta_{i+2} > \gamma_{i+2}$, ... and that every subsequent β_j is greater than the corresponding γ_j. The reader should not lose sight of the fact that all this formal verbiage is just another way of saying that we compare the digits of β and γ term by term (at corresponding places) going from left to right, and at the first place at which they differ the expression with the 1 at this place is the bigger of the two expressions. In particular, one sees immediately that

$$[1\,0\,1\,1\,0\,|\,0\,1\,1\,0\,1\,1\,\ldots.] > [1\,0\,1\,1\,0\,|\,0\,1\,1\,0\,0\,1\,\ldots.]$$

even if we do not know the missing digits.

We turn to the addition of $B = \beta A$ and $C = \gamma A$ where $\beta = [b_r \ldots b_0|b_{-1} \ldots]$ and $\gamma = [c_s \ldots c_0|c_{-1} \ldots]$. The sum $B + C = \beta A + \gamma A$ is a segment which can be expressed in terms of A, and in keeping with the notation when β and γ are integers or have finite dyadic expansions we denote this segment by $(\beta+\gamma)A$ — so $B + C = (\beta+\gamma)A$, and we must find the dyadic expansion of $(\beta+\gamma)$. For this, it is necessary to locate an infinite sequence of nested intervals associated with B + C (as expressed in terms of A); after all, given such a sequence of nested intervals, the dyadic expansion associated with it is precisely what we have denoted by $\beta + \gamma$. To carry this out, we make use of the nested intervals associated with B and C, which were described above in detail. According to the rules for adding inequalities, corresponding intervals in the nested sequences of B and C lead to the inequalities

$$(\beta_0+\gamma_0)A \leq B + C < (\beta_0+\gamma_0+2)A$$

$$(\beta_1+\gamma_1)A \leq B + C < (\beta_1+\gamma_1+1)A$$

$$(\beta_2+\gamma_2)A \leq B + C < (\beta_2+\gamma_2+\tfrac{1}{2})A$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$(\beta_t+\gamma_t)A \leq B + C < (\beta_t+\gamma_t+\frac{1}{2^{t-1}})A$$

$$(\beta_{t+1}+\gamma_{t+1})A \leq B + C < (\beta_{t+1}+\gamma_{t+1}+\frac{1}{2^t})A$$

$$\vdots \qquad \vdots \qquad \vdots$$

Note that the intervals in which B + C falls are of sizes

$$\begin{array}{l} 2A \\ A \\ \frac{1}{2}A \\ \vdots \\ \frac{1}{2^t}A \end{array}$$

Furthermore, because

$$\begin{array}{cccccccccccl}
B_0 & \leq & B_1 & \leq & B_2 & \leq & \cdots & \leq & B_t & \leq & B_{t+1} & \cdots \text{ and} \\
\gamma_0 & \leq & \gamma_1 & \leq & \gamma_2 & \leq & --- & \leq & \gamma_t & \leq & \gamma_{t+1} & --- \text{ it follows that} \\
\hline
B_0+\gamma_0 & \leq & B_1+\gamma_1 & \leq & B_2\gamma_2 & \leq & \cdots & \leq & B_t+\gamma_t & \leq & B_{t+1}+\gamma_{t+1} & \cdots
\end{array}$$

Consequently, in order to show that the intervals associated above with B + C are nested, it suffices to verify that

$$(\beta_0+\gamma_0+2) \geq (\beta_1+\gamma_1+1) \geq \ldots . \geq (\beta_t+\gamma_t+\frac{1}{2^{t-1}}) \geq (\beta_{t+1}+\gamma_{t+1}+\frac{1}{2^t}) \geq \cdots$$

To accomplish this we note that either $\beta_1 = \beta_0$ or $\beta_1 = \beta_0 + \frac{1}{2}$ (depending on whether the digit after the vertical stroke in β is a 0 or a 1), so that always $\beta_1 \leq \beta_0 + \frac{1}{2}$. In the same way, $\beta_2 \leq \beta_1 + \frac{1}{2^2}$, $\beta_3 \leq \beta_2 + \frac{1}{2^3}$ and, in general,

$$\beta_{t+1} \leq \beta_t + \frac{1}{2^{t+1}} \qquad t = 0, 1, 2, \ldots$$

Exactly the same procedure gives

$$\gamma_{t+1} \leq \gamma_t + \frac{1}{2^{t+1}} \qquad t = 0, 1, 2, \ldots$$

We have, therefore,

$$0 \leq \beta_{t+1} - \beta_t \leq \frac{1}{2^{t+1}} \qquad 0 \leq \gamma_{t+1} - \gamma_t \leq \frac{1}{2^{t+1}}$$

and adding gives

$$0 \leq \beta_{t+1} + \gamma_{t+1} - \beta_t - \gamma_t \leq \frac{1}{2^t}$$

which says that

$$\beta_t + \gamma_t - \beta_{t+1} - \gamma_{t+1} + \frac{1}{2^t} \geq 0 \qquad (*)$$

But we needed to prove that

$$\beta_t + \gamma_t + \frac{1}{2^{t-1}} \geq \beta_{t+1} + \gamma_{t+1} + \frac{1}{2^t}$$

which is just another way of writing (*).

All this says that we have a nested sequence of intervals which locates B + C; in particular, $(\beta_t + \gamma_t)A$ is the left hand endpoint of an interval of size $(\frac{1}{2^{t-1}})A$ in which B + C lies — so $(\beta_t + \gamma_t)A$ gives a very good approximation to $(\beta + \gamma)A = B + C$ as t gets bigger.

To illustrate: if $\beta = [1011|0110101....]$, $\gamma = [11000|101110011....]$ with the dots signifying that the missing digits are not given or simply omitted, then

$$\beta_0 + \gamma_0 = [100011|00...\text{(all 0)}...]$$

$$\beta_1 + \gamma_1 = [100011|100...\text{(all 0)}..]$$

$$\beta_2 + \gamma_2 = [100011|1100....\text{(all 0)}..]$$

$$\vdots$$

$$\beta_7 + \gamma_7 = [100100|001000100...\text{(all 0)}...]$$

Moreover, $[100100|001000100...]\ A = (\beta_7 + \gamma_7)A$ is $\leq B + C$, and is an approximation within $(\frac{1}{2^6})A$ of $B + C$ — that is, within $[0|00000100....]A$. In this case we see, therefore, that the expansion of $B + C$ starts with the digits $[100100|00100]$.

In general, we have no way of writing all the digits in the expansions of β and γ, so that there is no hope of writing all the digits of $\beta + \gamma$ - we must restrict ourselves to finite approximations (and this is what we always do when measuring in real life). However, there is one type of situation in which we have all the digits under control - namely, when the expansion eventually becomes periodic; in other words, the digits start to repeat themselves after a while, ad infinitum.

Consider, for example,

$$\beta = [110|\overset{\frown}{001}\ \overset{\frown}{001}\ \overset{\frown}{001}\ \ldots]$$

(where the notation $\overset{\frown}{001}$ indicates that the triplet 001 is repeated over and over) and

$$= [1010|\overset{\frown}{010}\ \overset{\frown}{010}\ \overset{\frown}{010}\ \ldots]$$

If we start to compute the approximations $\beta_t + \gamma_t$ to $\beta_9 + \gamma_9$, we see that among others

$$\beta_0 + \gamma_0 = [10000|00\ldots]$$

$$\beta_3 + \gamma_3 = [10000|0110000\ldots]$$

$$\beta_6 + \gamma_6 = [10000|011011]$$

$$\beta_9 + \gamma_9 = [10000|011011011]$$

Of course, $(\beta_9 + \gamma_9)A$ is $\leq B + C = (\beta + \gamma)A$ and the "error" is at most $\frac{1}{2^8}A$, so the expression for $\beta_9 + \gamma_9$ is the "correct" expression for $\beta + \gamma$ through 6 digits to the right of the vertical stroke.

By repetition of this process, it becomes clear that

$$\beta + \gamma = [10000|\overset{\frown}{011}\ \overset{\frown}{011}\ \overset{\frown}{011}\ \ldots.]$$

We may leave it to the reader to check, in detail, that

$$\beta + \beta = [1100|\overset{\frown}{010}\ \overset{\frown}{010}\ \overset{\frown}{010}\ \ldots]$$

$$\gamma + \gamma = [10100|\overset{\frown}{100}\ \overset{\frown}{100}\ \overset{\frown}{100}\ \ldots]$$

Things go smoothly in the preceding examples because the periods fit perfectly — what happens if they do not fit exactly. For example, suppose

$$\beta = [0|0\overset{\frown}{001}\ \overset{\frown}{001}\ \overset{\frown}{001}\ \ldots], \qquad \gamma = [0|\overset{\frown}{010}\ \overset{\frown}{010}\ \overset{\frown}{010}\ \ldots],$$

then it is fairly straightforward to convince oneself (from the approximations) that

$$\beta + \gamma = [0|0\ \overset{\frown}{101}\ \overset{\frown}{101}\ \overset{\frown}{101}\ \ldots]$$

- perhaps the quickest way to see this is to re-write γ in the form $\gamma = [0|0\ \overset{\frown}{100}\ \overset{\frown}{100}\ \overset{\frown}{100}\ \ldots]$.

What about the following situation?

$$\beta = [0|\overgroup{01}\ \overgroup{01}\ \overgroup{01}\ \ldots] , \qquad \gamma = [0|\ \overgroup{110}\ \overgroup{110}\ \overgroup{110}\ \ldots]$$

Taking a few approximations we have among others

$$\beta_2 + \gamma_2 = [1|00]$$

$$\beta_5 + \gamma_5 = [1|00101]$$

$$\beta_6 + \gamma_6 = [1|001011]$$

$$\beta_{12} + \gamma_{12} = [1|001100001011]$$

$$\beta_{18} + \gamma_{18} = [1|001100001100001011]$$

and it is not hard to convince oneself that

$$\beta + \gamma = [1|\overgroup{001100}\ \overgroup{001100}\ \ldots.]$$

The period of $\beta + \gamma$ is 6, essentially because we may re-write

$$\beta = [0|\overgroup{010101}\overgroup{010101}\ldots]$$

$$\gamma = [0|\overgroup{110110}\ \overgroup{110110}\ \ldots]$$

Appendix I.

How can we find the representation of an integer n in terms of powers of 2 ? Rather than begin this discussion in mathematical terms dealing with the integer n let us attack the problem from a more experimental nature.

Take a collection of 27 objects of the same type; for example, cards or pebbles will do. Group these 27 objects into piles of two elements, thus getting 13 piles of two and 1 element left over. Now taking the 13 two-element piles, we double these up, thus getting 6 piles of 4 elements each and one pile with two elements left over. So far, our 27 elements are distributed among 6 piles of 4 elements each, 1 pile with 2 elements and 1 pile with 1 element. Continuing in the same way, we double up the 6 piles of 4 elements and get 3 piles of 8 elements each. Again doubling up these 3 piles we arrive at 1 pile of 16 elements and 1 pile of 8. Obviously, we cannot do any further doubling up, and our set of 27 objects is "broken up" into 1 pile of 16, 1 pile of 8, 1 pile of 2, and 1 pile of 1. There can be no more concrete realization of the fact that

$$27 = 16 + 8 + 2 + 1$$

This straightforward mechanical procedure of doubling up clearly works for any positive integer n. It shows not only that n can be expressed as a sum of non-negative powers of 2, but also produces such an expression.

Exercise: Use the doubling-up method to express each of the following integers as a sum of powers of 2 ------- 78, 99, 129, 150, 250, 437, 500.

The method described above for finding the expansion of an integer in powers of 2 is informal, but surely thoroughly convincing. On the other hand, it requires objects for manipulation, and if n is large this is a matter of considerable inconvenience. Thus, it is not unimportant to give a formal, numerical explanation of our method, which shows how to find the expansion of any n (no matter how large) with minimal effort.

Consider any positive integer n; then upon division by 2 the remainder is either 0 or 1 (in fact, the remainder is 0 or 1 according as n is even or odd, respectively). In either case, we may write

$$n = 2n_0 + b_0 \qquad\qquad b_0 = 0 \text{ or } 1$$

and b_0 is the remainder upon division by 2. Note that this reflects exactly what was done with the concrete objects -- for example, if

$n = 27$, the 27 objects break up into $n_0 = 13$ piles of 2 objects each and there is $b_0 = 1$ object left over. Now, we may repeat the process for the positive integer n_0

$$n_0 = 2n_1 + b_1 \qquad\qquad b_1 = 0 \text{ or } 1$$

(Note that in the case $n = 27$ this represents the second step where the $n_0 = 13$ piles of 2 elements each with one two-element pile left over). Substituting this expression for n_0 in the expression for n we have

$$n = 2(2n_1 + b_1) + b_0 = n_1 \cdot 2^2 + b_1 \cdot 2 + b_0$$

Repeating this process, we have

$$n_1 = 2n_2 + b_2 \qquad\qquad b_2 = 0 \text{ or } 1$$

so that

$$n = (2n_2 + b_2)\, 2^2 + b_1 \cdot 2 + b_0$$

$$= n_2 \cdot 2^3 + b_2 \cdot 2^2 + b_1 \cdot 2 + b_0$$

and eventually we can divide no further and arrive at an expression

$$(\S) \qquad n = b_r 2^r + b_{r-1} 2^{r-1} + \dots + b_1 2 + b_0 \quad , \quad b_r = 1,$$

each $b_i = 0$ or 1. This is the canonical expression for n as a sum of powers of 2, and needless to say we abbreviate it by writing

$$b_r \, b_{r-1} \dots b_1 \, b_0$$

This is simply a finite sequence of zeros and ones (starting with a one, of course) which is a short-hand notation for $(\S)$, and it is often referred to as the "base 2" expansion of n.

In order to fix the procedure for finding the base 2 expansion of n in mind, it is useful to do some examples.

Suppose that n = 84, then

$$84 = (2)\ (42) + 0$$

which means that $n_0 = 42$, $b_0 = 0$; the next step gives

$$42 = (2)\ (21) + 0$$

so that $n_1 = 21$, $b_1 = 0$. Continuing this process, we get

$$21 = (2)(10) + 1$$

$$10 = (2)(5) + 0$$

$$5 = (2)(2) + 1$$

$$2 = 2 \cdot 1 + 0$$

How long does the process continue? Until we get an expression

$$n_{r-2} = 2n_{r-1} + b_{r-1} \qquad \text{with } n_{r-1} = 1$$

(this must always happen), whereupon we put $b_r = n_{r-1}$ and then have n as $b_r b_{r-1} \ldots b_1 b_0$. Thus, the remainders b_0, b_1, b_2, ..., b_{r-1} and the last $n_{r-1} = 1$ give precisely the digits of the base 2 expansion of n -- going from right to left. In particular, the base 2 expansion of 84, as read off from the list of remainders is

$$1010100$$

In similar fashion, for n = 179 we have

$$179 = (2)(89) + 1$$

$$89 = (2)(44) + 1$$

$$44 = (2)(22) + 0$$

$$22 = (2)(11) + 0$$

$$11 = (2)(5) + 1$$

$$5 + (2)(2) + 1$$

$$2 = (2)(1) + 0$$

and, therefore the base 2 expansion of 179 is precisely 10110011 which is short-hand notation for $179 = 128 + 32 + 16 + 2 + 1$

Appendix II

In this section we are interested in developing an economic method of adding in the base 2. By economic we mean both in notation and in computation.

As an example let us return to the text for an illustration;

11011 + 1000101

but instead of working horizontally let us work vertically after lining digits up carefully in columns. Labelling the columns at the top, for illustrative purposes, we are considering the addition

	2^7	2^6	2^5	2^4	2^3	2^2	2^1	$2^0 = 1$
				1	1	0	1	1
+		1	0	0	0	1	0	1
		1	0	1	1	1	1	2

Note that this, and everything we do here, is just a short-hand notation for the things done earlier. Now, replacing the 2 in the last column (or in any column) by a 1 in the next column to the left, we can keep re-writing -- and listing the various intermediate steps until we arrive at the final answer, we have

	2^6	2^5	2^4	2^3	2^2	2^1	$2^0 = 1$
	1	0	1	1	1	1	2
=	1	0	1	1	1	2	0
=	1	0	1	1	2	0	0
=	1	0	1	2	0	0	0
=	1	0	2	0	0	0	0
=	1	1	0	0	0	0	0

Thus, every line equals the preceding one, and the end-result is indeed 1100000.

In a sense, it is not the label at the top of a column that matters, but rather the relation between adjacent columns, and if we can keep the columns lined up accurately then the names of the columns and the "art-work" may be dispensed with. In this vein let us do 1011011 + 1001111 -- so

$$\begin{array}{c@{\qquad}ccccccc}
 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
+ & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\
\hline
 & 2 & 0 & 1 & 2 & 1 & 2 & 2
\end{array}$$

Now, there are several 2's to be adjusted, and it is important to appreciate the fact that these may be treated in any order. For example,

$$\begin{array}{c@{\qquad}cccccccc}
 & & 2 & 0 & 1 & 2 & 1 & 2 & 2 \\
= & & 2 & 0 & 2 & 0 & 1 & 2 & 2 \\
= & & 2 & 0 & 2 & 0 & 1 & 3 & 0 \\
= & 1 & 0 & 0 & 2 & 0 & 1 & 3 & 0 \\
= & 1 & 0 & 0 & 2 & 0 & 2 & 1 & 0 \\
= & 1 & 0 & 1 & 0 & 0 & 2 & 1 & 0 \\
= & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0
\end{array}$$

We conclude, therefore, that

$$\begin{array}{ccccccccc} & & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ + & & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}$$

Of course, not all the steps need to be written out in detail -- much of the work may be done mentally. In fact, if one proceeds through the columns in order from right to left, and keeps mental track of the "carrying" from one column to the next, then it is possible to write down the answer (term-by-term, and from right to left) without any intermediate steps. For example,

$$\begin{array}{ccccccccccccc} & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ + & & & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ \hline & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{array}$$

-- in the right column we have 1 + 1 = 2, so we write 0 and"carry" 1 to the next column. In this column, we now have 1 + 1 = 2, so we write 0 and carry 1. The third column has then 1 + 1 + 1 = 3 , so we write 1 and carry 1; the reader can easily complete the details.

Table of Contents

Laboratory Manual for Chapter I

1.1 - 1.4 Inequality 2

1.5 - 1.6 Equality and Its Properties 5

1.7 Addition and Its Properties 9

1.8 Multiplication by a Positive Integer. 11

1.10 - 12 Bisection of Segments 22

1.12 Dyadic Expansion of Segments 26

LABORATORY MANUAL FOR CHAPTER I

In this chapter we analyse the notion of measurement. We do so by performing a series of experiments. The materials consist of :

A simple balance

Assorted vials to hold liquid, beans, etc.

Assorted objects suitable for weighing, such as metal shavings, dried beans, etc.

Ruler

Compass

Some prepared plasticene sheets.

It is suggested that in the experiments involving the balances, two or three students work together to speed up the operations.

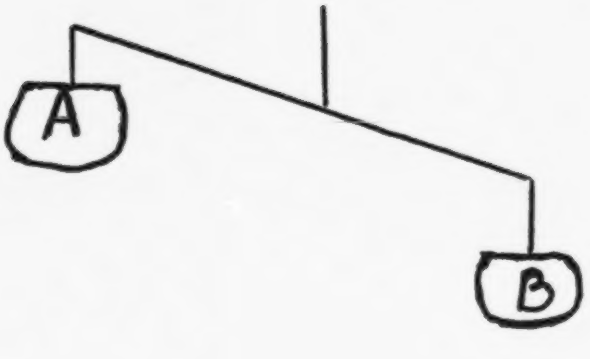
Be sure, in all weighing experiments using vials, that the same number of vials appear on both pans of the balance (adding empty vials if necessary). Otherwise, excess weight of the vials will render the experiment inaccurate.

1.1 - 1.4. Inequality

The most primitive notion underlying any situation in which some kind of measurement plays a role is that of inequality. An inequality is merely a way of making a comparison between two objects.

DEFINITION:

Inequality of two objects according to weight. Object A is put on one side of the balance, and object B is put on the other side. If the side containing A goes up while the side containing B goes down, we say that object A is lighter than B and write $A < B$. If side A is the one which goes down, we write $B < A$. The sign $<$ is to be read as less than.



$A < B$

Experiment 1

1. Fill two vials with unequal amounts of water. Label the one with less water A, and the one with the larger amount B. Compare these two vials on the balance.

2. Fill a third vial with a small amount of metal shavings. Have the volume of these metal shavings be less than the volume of

water in A. Label this vial C. Compare A with C. Write down your result using the symbol $<$. Do the same for B and C.

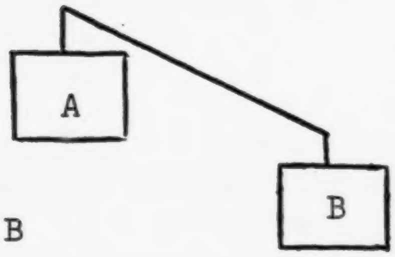
Retain the two vials A and B for use in the next experiment.

Discussion

Notice that when we compare two objects on the balance, we are really forgetting about all other relations between them other than there relative weight. A smaller volume of metal may weigh more than a larger volume of water. Our comparison introduces a certain amount of abstraction.

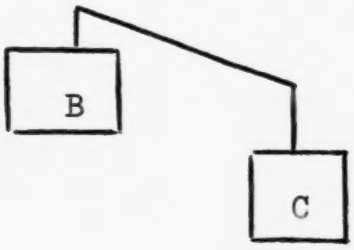
Experiment 2

1. On a balance demonstrate once again that $A < B$.
2. Now weigh out another vessel of water called C such that $B < C$.



$A < B$

2.



$B < C$

3. What will the balance look like when we put A on one side and C on the other? Indicate by drawing your prediction.
4. Check your predicted answer to 3 by comparing A with C.

Discussion

The first fundamental property of our "less than" relationship for weights or lengths is the transitive law. If A is lighter than B and B is lighter than C, then A is lighter than C; in symbols, if $A < B$ and $B < C$ then $A < C$. This rule, which is known as the transitive law, is so obvious that we often take it for granted.

In the text we point out (by Example) that for other kinds of comparisons, the transitive law need not hold. The transitive law for weights is thus really based on a collection of experimental facts. Whenever anyone has compared the weights of two objects A and B and found that $A \prec B$, compared B with C and found that $B \prec C$, then it has always turned out that a direct comparison of A with C showed that $A \prec C$. This has happened so consistently that we believe it to be true in all cases. As with all physical laws we then use the transitive law as a basis for deduction. If we are informed that $A \prec B$ and $B \prec C$, we conclude that $A \prec C$ without directly comparing A with C.

Experiment 3

Directions

1. Make an alphabetical list of all students in the class.
2. The first person on the list will weigh out a rather small weight and mark it A.
3. This first person will pass this weight A to the next person on the list.
4. The second person should make a weight and mark it B such that $A \prec B$.
5. Pass this weight B to the third person.
6. The third person will make a weight C such that $B \prec C$.
7. Continue in this manner until the last person makes a weight.
8. What is the relation between this last weight and weight A?, weight B?, weight C?

Discussion

We could have predicted the outcomes of this experiment by repeated application of the transitive law: If $A < B$ and $B < C$ then $A \prec C$. If $C \prec D$ then since $A \prec C$ and $C \prec D$ we deduce that $A < D$, and so on.

The actual experiment is performed in order to contrast it with Experiment 7.

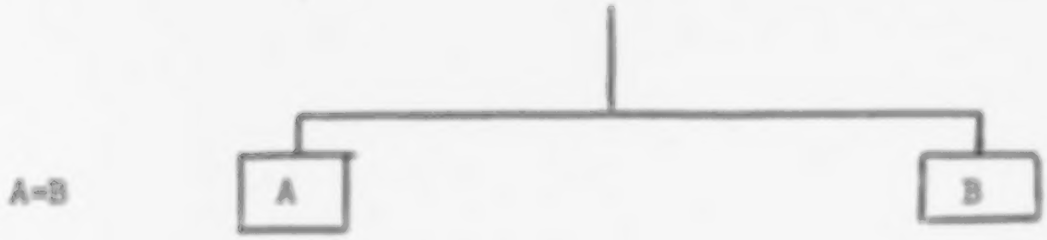
TRANSITIVE PROPERTY.

If A, B and C are any three weights the following statement is true:

if $A < B$ and $B < C$ then $A < C$.

1.5 - 1.6 Equality and its Properties

If objects A and B are placed on different sides of a balance and neither side goes down, that is, if the two sides balance, then we say that object A is equal in weight to object B. For simplicity, we then write A=B, although such a notation obviously leaves much to be desired. Thus A=B does not mean that A is B. It only means that A and B balance each other on the scale.



In checking for equality be sure to interchanve A with B and the balance.

To be sure of equality, remove A and B from the balance and then replace them, (perhaps on opposite sides). This is to help avoid the interference from the friction of the balance.

Experiment 4

1. Put some beans in one vial. Mark it A. Fill vial B with water so that A=B.

Discussion

This experiment shows that (with some difficulty) we can reproduce any given weight. That is, starting with any object A we can find another object B that weighs the same.

Experiment 5

1. Pour some water into a vial and mark it C.
2. Fill vial D with metal bolts so that $D=C$. Can this be done?

Discussion

An essential property of weight as opposed to number is that it is not discrete. We may not be able to reproduce a given weight by a number of multiples of some other weight.

The transitive law for equality

Experiment 6

1. Choose an object A and weigh out a vial B of water equal in weight to A.
2. Weigh out an object C equal in weight to B.
3. Compare C with A.

Experiment 7 Repeat Experiment 3 for equality. That is,

1. Make a list of the students in the class
2. Let the first person on the list pick an object A, reproduce an object B equal in weight to A and pass B to the second person on the list, returning A.
3. The second person then carefully weighs out C equal in weight to B and passes C to the third person on the list and returns B.
4. Continue in this way to the last person on the list.
5. What do you expect the relation of the last object and A to be?
6. Compare the last object with A.

Discussion

The first fundamental observation about the relationship of equality is again the validity of the transitive law. That is, if $A=B$ and $B=C$ then $A=C$. However, in contrast to the transitive law for inequality, the transitive law for equality is frequently an idealization from experience rather than something that always holds true in practice. Thus, if we have objects A, B, C, D, E with A=B, B=C, C=D and D=E then standard rules of reasoning lead to the conclusion that A=E. Unfortunately, experiment 7 shows that in practice this assertion often breaks down. We tend to think of the transitive law as "logically obvious."

The reason for this apparent contradiction of experience with the rules of logic is, of course, the inaccuracy of our balance. There is a certain amount of experimental error involved in each weighing. Thus although A and B balance on our rough balance, they are probably not really equal in weight, that is, the use of a more delicate and accurate balance could show this. Now, such errors can accumulate sufficiently so that they do indeed show up even on our rough balance; this is why the experiment led to an unexpected result. Unfortunately, this accumulation of error is unavoidable. If we were to use extremely delicate balances, the same trouble would arise, because, after all, no balance is truly perfect.

It may be remarked that if Experiment 7 is repeated a number of times, it will turn out that sometimes the end product is lighter than A, sometimes it equals A, and sometimes it is heavier than A. If things work reasonable well, the end product turns out to be lighter than A or heavier than A with equal frequency. This indicates that the break-down of the transitive law for equality does not reflect something that is fundamentally missing from the relation -- rather, it is due simply to accumulation of experimental error. The cases in which the end product is equal to A in weight occur precisely when the various experimental errors cancel each other -- some students may produce weights which are too heavy while others may produce weights which are too light.

In summary, the transitive law for equality is a rule which we regard as holding in an ideal situation. According to our viewpoint, the equality represented by a balance is merely a crude approximation to the ideal equality that we would expect to hold for an ideal balance.

If A weighs the same as B we shall write

$$w(A) = w(B).$$

The idea of this notation is that we can replace the relation between the objects A and B by an assertion concerning an "abstract property" of A and of B. Instead of saying that A and B balance out the scale, we say that the "weight of A" equals the "weight of B." We have attached to each real object A an abstract property, $w(A)$, which is called its weight. (It is important to observe that $w(A)$ is not a number). Two objects "have the same weight" if they balance. The general way in which abstract properties are attached to real objects is via the notion of equivalence relation. This is discussed in the text. In terms of the notion of equivalence class, we can say that the weight of A is the equivalence class to which A belongs.

Experiment 8

1. Choose objects A and B with $A < B$.
2. Weigh out objects C and D such that $w(A)=w(C)$ and $w(B)=w(D)$.
3. Compare C with D.

Discussion

The experiment shows that if $A < B$ and $w(A)=w(C)$ and $w(B)=w(D)$ then $C < D$. Thus in the inequality $A < B$ between two real objects A and B, we could replace A by any other object of the same weight and replace B by **any** other object weighing the same as B and the inequality will still hold. This shows that we really have an inequality between the weight of A and the weight of B and we can write

$$w(A) < w(B) \quad .$$

This is now an inequality relating the abstract concepts $w(A)$ and $w(B)$. It says choose any object whose weight is $w(A)$ and you will find that it weighs less than **any** object whose weight is $w(B)$.

Of course, the transitive law holds for the notion of inequality of two weights:

$$\text{if} \quad w(A) < w(B) \quad \text{and} \quad w(B) < w(E) \quad \text{then} \quad w(A) < w(E).$$

1.7 Addition and its properties.

Consider any two objects A and B, and combine them by lumping them together into a single pile. This pile may be viewed as a new object which we denote by A+B. From the point of view of our balance, A+B means simply that both A and B are placed together on the same side of the balance. Since it clearly does not matter in what order A and B are placed on the same side of the balance, there is no way to distinguish between A+B and B+A; therefore, we must view A+B and B+A as the same object -- that is, A+B = B+A.

Experiment 9

1. Choose objects A and B. Get objects A' and B' such that $w(A')=w(A)$ and $w(B')=w(B)$.

2. Compare $A+B$ with $A'+B'$.

Discussion

If $w(A)=w(A')$ and $w(B)=w(B')$ then $w(A+B)=w(A'+B')$. This shows that $w(A+B)$ depends only on $w(A)$ and $w(B)$ and not on the specific objects A and B. It therefore makes sense to write $w(A)+w(B)$ where it is understood that we are making the definition

$$w(A) + w(B) = w(A+B) .$$

This definition says: we add the weights $w(A)$ and $w(B)$ as follows: pick any object A of weight $w(A)$ and any object B of weight $w(B)$ and bring them together to get $A+B$. Then we define $w(A) + w(B)$ to be $w(A+B)$. This definition makes sense because of the outcome of Experiment 9. If we chose some other weight A' instead of A and some other weight B' instead of B, then we would end up with the same weight -- $w(A+B)=w(A'+B')$.

This operation of addition provides a crucial step towards our goal of assigning numbers to abstract properties such as weights. With this objective in mind we need, first of all, to observe that the usual rules for addition of numbers are valid for this operation of addition of weights. We also need to understand how this relation of addition interacts with the relation of inequality between weights.

Experiment 10

1. Select three objects A, B, and C.
2. Construct an object D such that $w(D)=w(A+B)$.
3. Construct an object E such that $w(E)=w(B+C)$.
4. Compare $w(A)+w(E)$ with $w(D)+w(C)$.

Discussion

Experiment 10 illustrates the associative property:

$$\big(w(A)+w(B)\big) + w(C) = w(A)+\big(w(B)+w(C)\big).$$

1-8. Multiplication by a positive integer

Warning! Change in convention! From now on we are going to make a basic change in our convention. We are going to use the symbol A to denote the weight of an object instead of $w(A)$. We shall also use the symbol A to denote any object having the weight $w(A)$. So, we will say "reproduce weight A" instead of using the more cumbersome (but more precise) language "construct an object A' such that $w(A)=w(A')$." We will say "form A+B" when we mean "construct an object C such that $w(C)=w(A+B)$." We will tolerate this slight misuse (or imprecision) of language in order to have a little more smoothness of expression.

From the preceding section, we know how to add weights; thus, for any weight A we may define $2A = A+A$, $3A = A+A+A$, and, in general, for any positive integer n, $nA = A+A+\ldots.+A$, where there are n copies of A in the sum on the right. Note that for $n=1$, the definition says that $1A = A$. This operation, in which we take a positive integer and a weight and "combine" them to get a weight will be called "multiplication by a positive integer."

An integer times a weight is another weight. This operation is quite distinct from the product of two integers. (It makes no sense to multiply two weights nor does it make any sense to say $A \cdot n$).

There are several natural and important properties of this operation. From the associative law for addition it follows that if m and n are positive integers and A is an arbitrary weight then

$$(m+n)\, A = mA + nA$$

and

$$(mn)\, A = m\,(nA)$$

Note that in the first of these equations the addition on the left side is for integers, while on the right side it is addition of weights. In addition it follows from the associative and commutative laws for addition that if n is any positive integer, we can illustrate the first of these equations by the following:

Experiment 11

1. Choose a weight A
2. Form 2A and set it aside
3. Form 3A and set it aside
4. Form 5A (by reproducing A five times).
5. Compare 2A+3A with 5A.

Since $5=2+3$ we can rewrite the result of step 5 as $2a+3a=(2+3)A$

The equation $(m+n)A=mA+nA$ is called the first distributive law.

The second distributive law.

Experiment 12

1. Select any two random weights, A and B, by pouring two arbitrary amounts of water into two different cylinders.
2. Using a balance produce the following weights and designate the weight on the container:
    a. A+B
    b. Reproduce A+B.
    c. 2 x (A+B)
    d. 2 x A.
    e. 2 x B.

3. Compare on a balance the weight 2 x (A+B) with the weight 2 A + 2 B.

Discussion

The distributive law says that for any integer n and any weights A and B, if we form A+B and then multiply by n it is the same as multiplying A by n and B by n and then adding; symbolically

$$n\,(A+B) = nA + nB\ .$$

INEQUALITIES

If A and B are weights with $A < B$ and if C is any other weight then

$$A + C < B + C.$$

If you like, you can devise and carry out the experiment which verifies this.

If $A < B$ and $C < D$ then

$$A + C < B + C$$

while

$$B + C < B + D$$

so, by the transitive law

$$A + C < B + D.$$

If $A < B$ then (if we let C=A and D=B in the previous inquality)

$$A + A < B + B$$

or, in other words

$$2A < 2B\ .$$

For the same reason

$$3A < 3B$$

and, in general,

$$mA < mB$$

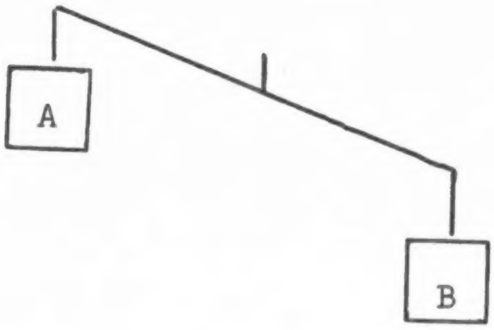
for any positive integer m.

If $A < B$ we do not know how $2A$ compares with B.

Experiment 13

1. Choose a weight A.
2. Demonstrate equality in the balance to get a number of equal weights.

3. Put several copies of A aside for use.
4. Measure out a weight B such that $A < B$.



5. On the same side as A put on more A's until $B < A+A+A+..+A$.

$$B < mA$$

6. What is the smallest m that works? $m =$ ——

Discussion

We started with two weights A and B such that $A < B$. We found an integer m such that

$$B < mA$$

but such that B is not $< (m-1)A$. We can say that either

$$(m-1)A < B$$

or $$(m-1)A = B$$

As a shorthand notation, we shall write

$$(m-1)A \leq B$$

which is to be read as $(m-1)A$ is "less than or equal" to B. We thus have

$$(m-1)A \leq B$$

and

$$B < mA \ .$$

We shall frequently combine these two inequalities by simply writing

$$(m-1)A \leq B < mA \ .$$

In other words, we know that B is at least as large as $(m-1)A$ but definitely smaller than mA.

There is, of course, at most one integer A that works. This integer m gives us a better idea of how B compares with A . There is, for example, much more information in the assertion

$$4A \leq B < 5A$$

than in the assertion $A < B$. Starting with A and B, can we always find a suitable m ? Is it possible that A is so small compared to B that no matter how many copies of A we add to itself we never exceed B ?

Experiment 14

1. This time our A is to be a drop of water measured from the standard eye-dropper in your kit.
2. Put a weight B on the balance and a container with one drop on the other side.

$A < B$

3. With the eye dropper add enough A's until the balance looks like fig. 4.g (2).

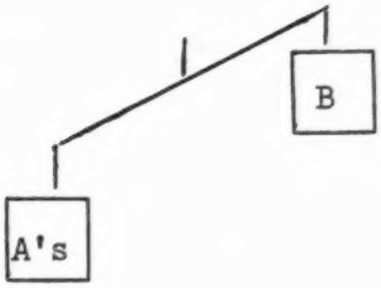


Fig. 4 g (2)

4. How many drops did you need?

The Archimedean principle asserts that given any weights A and B, $A<B$ there always will be some integer m such that

$$B < mA .$$

Experiment 15

1. Choose a weight A and reproduce several copies of A for use in this and the following experiment.
2. Choose weights B and C significantly different from A .
3. Find m such that

$$mA \leq B < (m+1)A$$

(Notice the shift in notation from the last experiment. If $4A \leq B < 5A$ then $m=4$!)

4. Find an integer n such that

$$nA \leq C < (n+1)A$$

5. Construct B+C. Find R such that

$$RA \leq B+C < (R+1)A.$$

Experiment 16

Use the weight A of the previous experiment

1. Find a weight E such that

$$2A < E < 3A$$

2. Find a weight F such that

$$5A < F < 6A$$

3. Construct E + F. Find the integer p such that

$$pA \leq E + F < (p+1)A.$$

p= ——.

Retain A, E and F for the next experiment.

Discussion

If $2A < E$ and $5A < F$ then the law of addition of inequalities tells us that

$$2A + 5A < E + F$$

so, in other words

$$7A < E + F .$$

Similarly

$$E < 3A \quad \text{and} \quad F < 6A$$

tell us that

$$E + F < 3A+6A = 9A .$$

So we know, in advance that

$$7A < E + F$$

and

$$E + F < 9A.$$

So, in Experiment 16, we could have predicted in advance that $p=7$ or $p=8$. We can't tell, in advance, which of these is correct.

3

But we have made a first step towards relating the addition of weights to the addition of numbers. The next step is to try to refine the information relating B to A by comparing B with multiples of $\frac{1}{2}A$. For this we must construct $\frac{1}{2}A$.

Experiment 17

1. Find a weight G such that $G+G=A$. We call this weight $\frac{1}{2}A$.

2. Using the weight E of the last experiment, decide which of the following assertions is true

$$2A \leq E < 2A+\tfrac{1}{2}A \quad \text{or}$$

$$2A+\tfrac{1}{2}A \leq E < 3A$$

3. Decide which of the following assertions is true

$$5A \leq F < 5A+\tfrac{1}{2}A \quad \text{or}$$

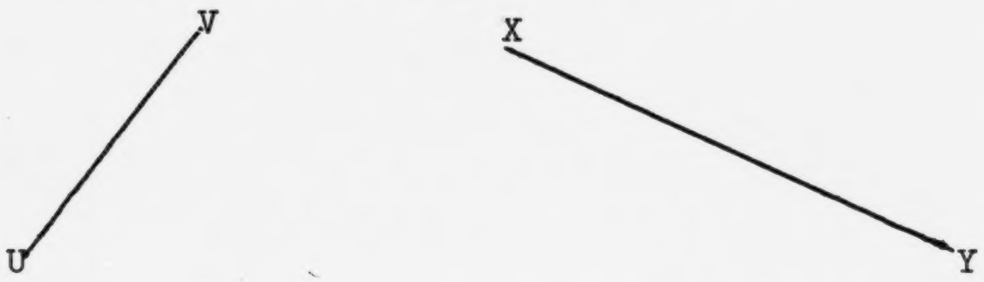
$$5A + \tfrac{1}{2}A \leq F < 6A$$

4. From the true assertions in 2 and 3, can you deduce an assertion relating E+F to A which is more refined than $7A \leq E+F < 9A$. What is this more refined assertion?

To continue our analysis, we would want to have $\frac{1}{4}A$ at our disposal. We would find it by subdividing $\frac{1}{2}A$ into two equal parts.

Since dividing a weight in two is a difficult and tedious process, we will now switch from our study of weights to a study of length of line segments. We should bear in mind that the experiments we will be performing with segments could theoretically be carried out with weights.

Let UV and XY be two line segments.



We compare their lengths as follows: Open the compass so that one point lies on U and the other lies on V. Place the compass with this opening with one point at X. If the other point does not reach as far as Y, we say that $\overline{UV}$ is shorter than XY and write

$$l(\overline{UV}) < l(\overline{XY}) .$$

If the other end of the compass fits exactly at Y, we write

$$l(\overline{UV}) = l(\overline{XY}) .$$

We can check that the inequality involving length satisfies the transitive law. We can also check that the relation

$$l(\overline{UV}) = l(\overline{XY})$$

is an equivalence relation. We can therefore study the corresponding abstract property known as length. We shall denote segment lengths by letters a,b,c, etc.

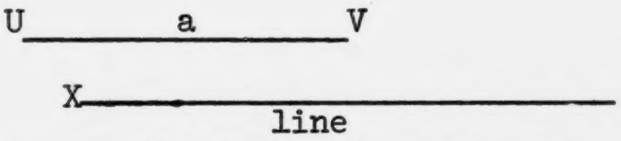
Before proceeding, we recall a number of constructions from plane geometry.

Basic Geometric Constructions using a compass and an unmarked straight-edge

Construction No. 1

Reproducing a line segment on the given line.

1. Given segment a.



2. Put the compass point on the left end-point U of a and open the compass until it spans segment a.

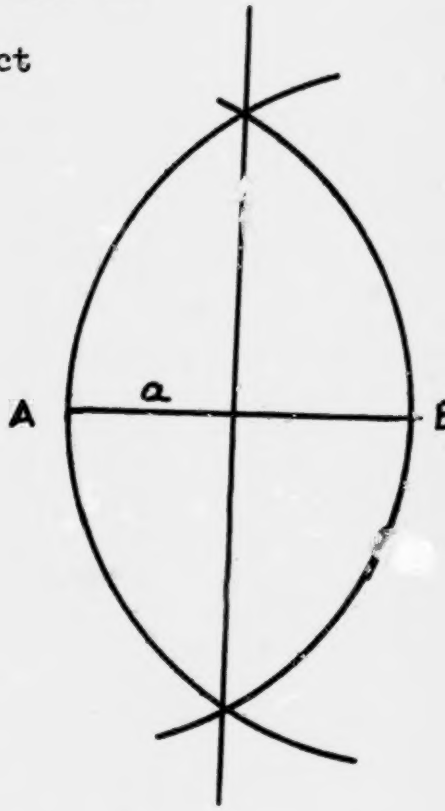
3. Keeping this opening, put the compass point a point X on the line and strike an arc through the line.

4. This is reproducing segment a on the line.

Construction No. 2

Bisecting a line segment.

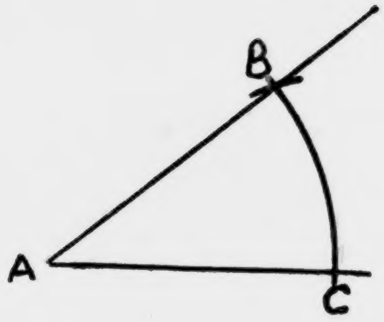
1. Given a line segment a with end points A and B.
2. With A as center and $\overline{AB}$ as radius construct a circle.
3. With B as center and $\overline{BA}$ as radius construct a circle.
4. Draw a segment using for endpoints the intersections of the above constructed circles.
5. This newley constructed segment is the bisector of $\overline{AB}$, and is also perpendicular to AB.

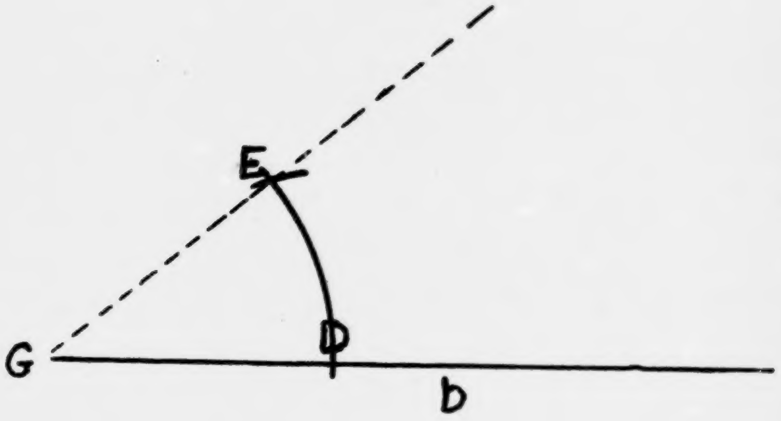


Construction No. 3

Duplicate an angle

1. Given an angle A and a segment b





2. Using A as center strike an arc intersecting with A at B and C.

3. Maintain same radius and strike an arc with left endpoint (G) of b as center and intersecting b at D.

4. Transfer segment $\overline{BC}$ to point D such that the arc's intersect at E.

5. Draw $\overline{GE}$.

6. $\measuredangle BAC = \measuredangle EGD$

Construction No. 4

Constructing a line parallel to a given line through a point.

1. Given a line l_1 and a point D not on the line.

2. Choose a point C on l_1 and draw the line through C and D.

3. Let A and B be points on l_1.

4. At D and on the same side of $\overline{CD}$ as B, construct an angle CDF congruent to ACD.

5. Draw the line through D and F. This line is parallel to l_1.

Construction No. 5

Divide a line segment into n congruent parts.

1. Given segment a to be divided into n con- n congruent parts. The endpoints of segments a are A and D.

2. From the left endpoint of a draw a line l

3. On l_1 lay off n congruent segments,

$$\overline{AC_1} = \overline{C_1C_2} = \overline{C_2C_3} = \dots = \overline{C_{n-1}\ C_n}$$

4. Through D draw the line parallel to l_1.

5. On this line, starting from D lay of n segments congruent to AC . Call D=Dn, the next point Dn-1 and so on.

6. Join Cn to Dn, Cn-1 to Dn-1, etc.

7. The intersections of these lines with a subdivide a into n equal parts.

Construction No. 6

Addition of two segments

1. Given two segments a and b.

2. Draw a line 1

3. Reproduce a on 1 as in Construction 1.

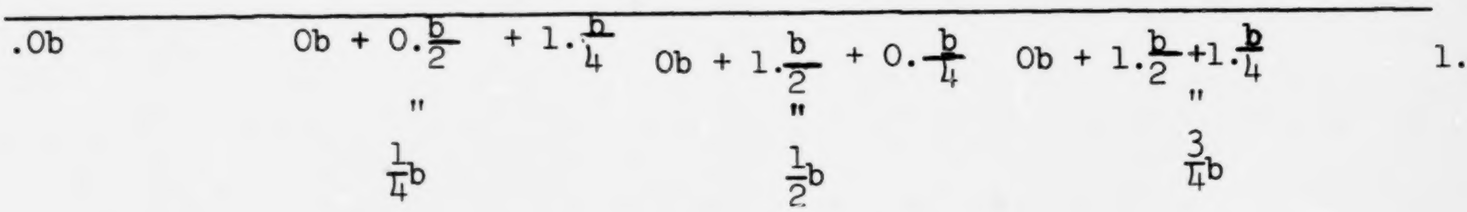
4. Starting at right endpoint of a, reproduce b on 1.

5. Segment (a+b) begins at left endpoint of a and ends at right endpoint of b.

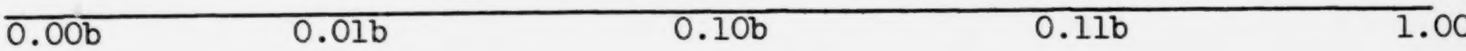
1.10 - 12 Bisection of Segments

Let us bisect an arbitrary segment b and designate the length or each bisected segment as $\frac{b}{2}$. Now bisect the resultant segment $\frac{b}{2}$ and designate its length as $\frac{1}{2}\left(\frac{b}{2}\right) = \frac{b}{4}$. Repeat the bisecting for the segment $\frac{b}{4}$ and designate its length as $\frac{b}{8}$. In general, if this bisection is repeated n times on the resultant segment the length of th the final segment will be designated as $\frac{b}{2^n}$.

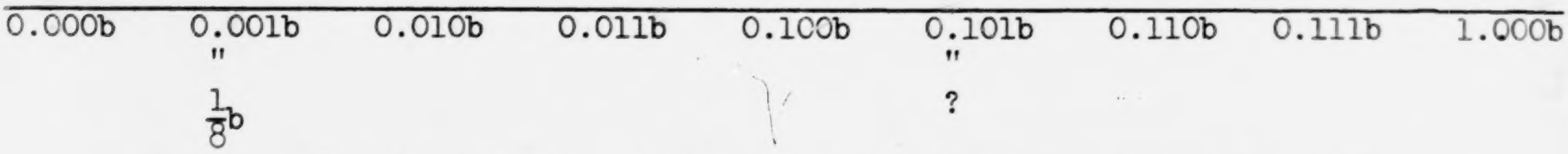
These lengths, after the bisection has been repeated twice, can be represented by means of the following diagram:


.0b
$0b + 0.\frac{b}{2} + 1.\frac{b}{4}$
"
$\frac{1}{4}b$
$0b + 1.\frac{b}{2} + 0.\frac{b}{4}$
"
$\frac{1}{2}b$
$0b + 1.\frac{b}{2} + 1.\frac{b}{4}$
"
$\frac{3}{4}b$
1.

or using a notation that is less cumbersome we have



After the bisection has been repeated three times the lengths of the segments may be represented as follows:


0.000b
0.001b
"
$\frac{1}{8}b$
0.010b
0.011b
0.100b
0.101b
"
?
0.110b
0.111b
1.000b

In general then, if the bisection has been repeated n times, so that the length of the segment is $\frac{b}{2^n}$ then the binary expansion of $\frac{b}{2^n}$ is represented as $\underbrace{0.000\ldots1b}_{n \text{ digits}}$

The following table shows the representation of the length of each segment as the number of bisections of segment b increases.

Number of bisections of Segment 6	Length of each segment	binary expansion
1	$\frac{b}{2}$	0.1b
2	$\frac{b}{2^2}$	0.01b
3	$\frac{b}{2^3}$	0.001b
4	$\frac{b}{2^4}$	0.0001b
:	:	:
n-1		$\underbrace{0.000\ldots1b}_{n-1 \text{ digits}}$
n	$\frac{b}{2^n}$	$\underbrace{0.000\ldots1b}_{n \text{ digits}}$

As the number of bisections increase, it follows that the length of each segment decreases. This is reflected in the binary expansion of the segments. Thus

$$\frac{b}{2} = 0.1b, \quad \frac{b}{4} = 0.01b, \quad \frac{b}{8} = 0.001b, \quad \ldots, \frac{b}{2^n} = 0.\underbrace{000\ldots1}_{n \text{ digits}}$$

Hence for bisection of segments the larger of two segments is the segment which possesses a digit 1 in the left-most position.

By referring to the line diagram and the table for the bisected line segments upon adding the segment of length $\frac{b}{2^2}$ to the segment of length $\frac{b}{2^2}$ we have

$$\begin{array}{rrcl} & \frac{b}{2^2} & = & 0.01b \\ + & \frac{b}{2^2} & = & 0.01b \\ \hline & \frac{b}{2} & = & 0.1b \end{array}$$

This may be expressed as $2 \times \left(\frac{b}{2^2}\right) = 2 \times (0.01b) = 0.1b$.

In general $m \times \left(\frac{b}{2^N}\right)$ designates $\frac{b}{2^N} + \frac{b}{2^N} + \ldots + \frac{b}{2^N}$ (m times).

The following addition of segments are similarly true:

$$\begin{array}{rrcl} & \frac{b}{2^3} & = & 0.001b \\ + & \frac{b}{2^3} & = & 0.001b \\ \hline & \frac{b}{2^2} & = & 0.010b \end{array} \qquad\qquad \begin{array}{rcl} \frac{b}{2^3} & = & 0.001b \\ \frac{b}{2^2} & = & 0.010b \\ \hline \frac{3 \times b}{2^3} & = & 0.011b \end{array}$$

From these examples the following addition facts must be true:

$$\begin{aligned} 0 + 0 &= 0 \\ 0 + 1 &= 1 \\ 1 + 0 &= 1 \\ 1 + 1 &= 10 \end{aligned}$$

Exercise : Find the sum of the two segments $\frac{b}{2^3}$ and $\frac{b}{2^4}$ both by means of the line diagram and by means of adding their binary expansions.

It was previously shown that $2 \times (0.01b) = 0.1b$. Multiplying by two, or equivalently doubling the size of the segment, results then in shifting the radix point one place to the right. Multiplying by four, that is, multiplying by two twice, results then in shifting the radix point two places to the right.

Exercise : What is the result of multiplying a binary number by 2^n?

Now let us examine multiplication more closely for the purpose of developing a multiplication algorithm. Consider the product $3 \times (0.01b)$

$$\begin{aligned} 3 \times (0.01b) &= (2+1) \times (0.01b) \\ &= 2 \times (0.01b) + 1 \times (0.01b) \\ &= 0.1b + 0.01b \\ 3 \times (0.01b) &= 0.11b \end{aligned}$$

This example may be abbreviated by means of the following algorithm:

$$3 \times (0.01b) = \begin{array}{r} .01b \\ \underline{11} \\ .01b \\ \underline{0.1\ b} \\ 0.11b \end{array}$$

For multiplication then the following multiplication facts must be true:

$$\begin{array}{l} 0 \times 0 = 0 \\ 0 \times 1 = 0 \\ 1 \times 0 = 0 \\ 1 \times 1 = 1 \end{array}$$

Exercise: Find the product of 5 x 0.001b both by means of the line diagram and by means of the multiplication algorithm.

Now consider the addition fact previously established: 0.001b + 0.010b = 0.011b. This addition fact is equivalent, by the definition of subtraction, to the statement 0.011b - 0.010b = 0.001b, that is the answer to 0.011b - 0.010b is the number which when added to 0.010b yields 0.011b. Thus:

$$-\begin{array}{r} 0.011b \\ \underline{0.010b} \\ 0.001b \end{array}$$

Exercise: Find the difference of the two segments 0.111b and 0.101b both by means of the line diagram and by means of subtracting their binary expansions.

Now consider the multiplication fact previously established:

$$3 \times (0.01b) = 0.11b$$

3 x (0.01b) = 0.11b is equivalent, by the definition of division, to the statement 0.11b ÷ 0.01b is the number of times 0.01b can be subtracted from 0.11b till 0.b is left. Thus 0.11b ÷ 0.01b may be obtained as follows:

$$\begin{array}{rrl} - & 0.11b & 1 \\ & \underline{0.01b} & \\ - & 0.10b & 1 \\ & \underline{0.01b} & \\ - & 0.01b & 1 \\ & \underline{0.01b} & \\ & & \textcircled{3} \end{array}$$

This repeated subtraction approach may be abbreviated by the following algorithm:

$$\begin{array}{rl} & \underline{11} = 1(2) + 1(1) = 3 \\ 0.01b & \big)\,0.11b \\ & \underline{.10b} \\ & .01b \\ & \underline{.01b} \end{array}$$

Exercise: Find the quotient 0.011b ÷ 0.001b by the process of repeated subtraction and by the division algorithm.

1.12 Dyadic Expansion of Segments

Experiment 18

In this next series of experiments we are going to investigate the nature of a real number as a sequence of "nested inequalities."

The method by which this investigation is to be carried out is to compare an arbitrary segment in terms of a given or chosen segment. The comparison is to be made using the dyadic expansions as explained in the previous pages.

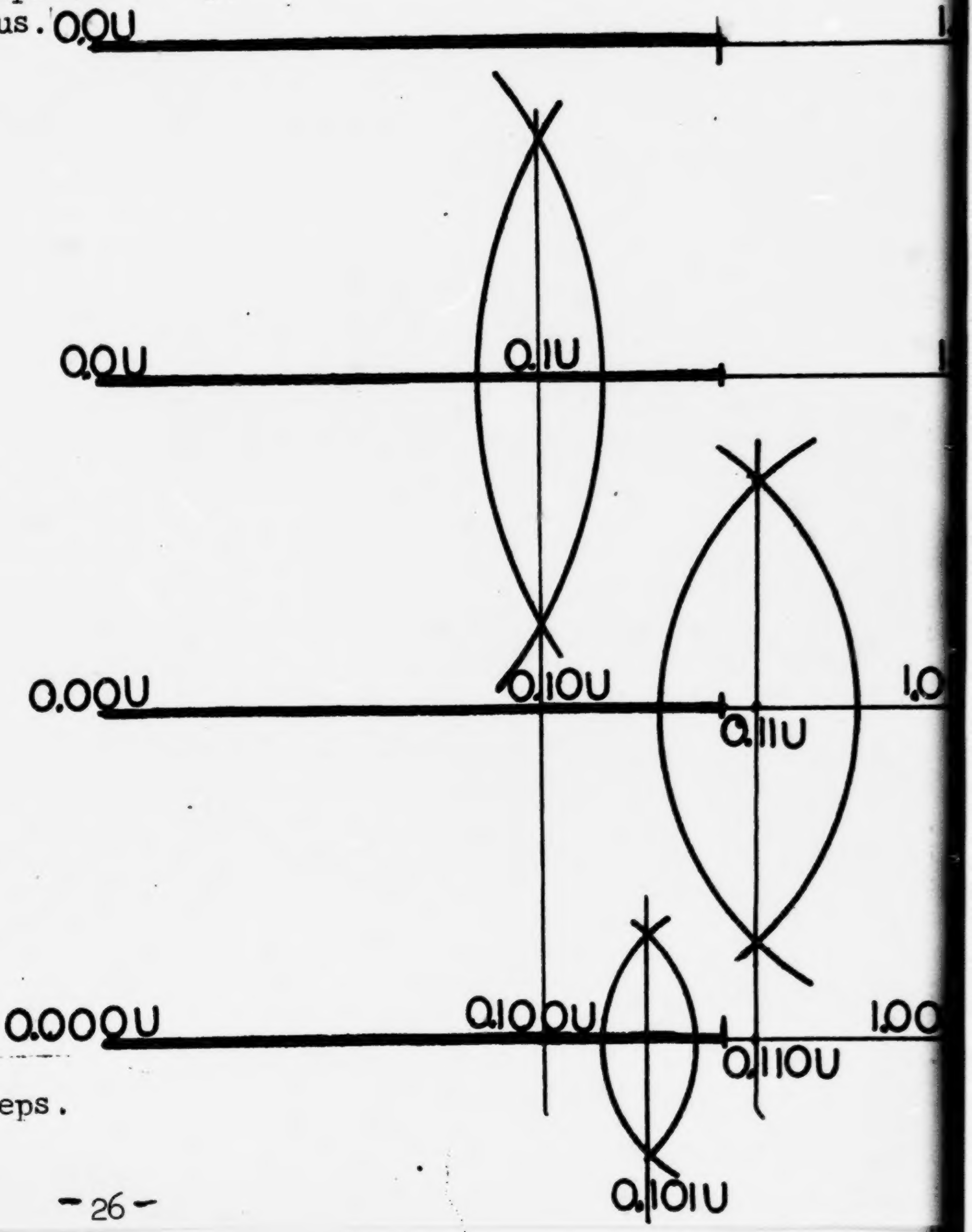
As an aid to your work the following example should be noted.

Given an arbitrary segment a and a unit segment u.

No. 4 is obviously a better approximation of a in terms of our unit u than is No. 1, but there is still room for improvement if our tools permit us.

$0.0u < a < 1.0u$
$0.1u < a < 1.0u$
$0.10u < a < 0.11u$
$0.101u < a < 0.110u$

The illustration at the right indicates the successive steps you will be taking but of course you will do this work on one line.



Carry out the next two steps.

Experiment 19

Directions

1. Choose a segment equal in length to the width of your four fingers. Call it a.
2. Construct a unit segment equal to three times the length of a. Call it b.
3. Find the binary expansion of a in terms of b.
4. Use the accompanying table as a guide.

$0.0b < a < 1.0b$
$0.0b < a < 0.1b$
$< a <$
$< a <$
$< a <$
$< a <$
$< a <$

What is the general pattern?

If you do not see the general pattern, then repeat the experiment with the segment a two or three times as large as the one you are now using and carry it out to eight or more steps.

Homework assignment

Experiment 19c

Directions

1. Choose two segments c and d such that $d = 7c$.
2. Find the binary expansion of c in terms of d.
3. It is important, since we want a fair degree of accuracy, that a rather large segment d should be chosen and that the constructions be as accurate as possible.
4. For convenience and uniformity, use the accompanying table.

$0.0d < c < 1.0d$
$0.0d < c < 0.1d$
$0.00d < c < 0.01d$
$< c <$
$< c <$
$< c <$
$< c <$

What is the general pattern? What will be the answer if we can get three more stages of accuracy? Will the procedure ever terminate whith exact equality?

Experiment 20

Addition of Segments

Directions

1. Choose any fairly large segment u as the unit of comparison. Keep u for the next few experiments.
2. Choose segments a and c and then construct a segment of length (a+c).
3. Find the binary expansions of a and c in terms of u to five places.
4. Find the binary expansion of (a+c) to five places.
5. Compare the binary expansion of a plus the binary expansion of c with the binary expansion of (a+c).

   Show your computations here.

   What are your conclusions?

Experiment 21

Directions

1. Construct an isosceles right triangle with a leg equal to a, where a is a fairly large segment.
2. Find the binary expansion of the hypotenuse in terms of a, to five places.
3. Recall the rule for multiplication of dyadic expansions (p.90-95) of the text. Multiply the binary expansion obtained in stage 2 by itself. What is the answer? What do you think the answer should be if the binary expansion were carried out to ten places?

Suppose that the expansion of segments m, n, p, q and r in terms of segment a are as follows:

$$0.100000a \leq m < 0.100001a$$

$$0.011000a \leq n < 0.011001a$$

$$0.101100a \leq p < 0.101101a$$

$$0.010101a \leq q < 0.010110a$$

$$0.001010a \leq r < 0.001011a$$

As these dyadic expansions are represented in terms of the binary expansions of segment a it follows that we can use the binary expansions to compare segments. The larger of two segments is the segment which possesses a digit 1 in the left most position; if both segments possess a digit 1 in the same left position the same comparison is made for each digital position to the right until the two segments possess a different digit.

Exercise: Using the preceding definition order by magnitude segments m, n, p, q, and r.

Experiement 22

Directions

1. Choose a fairly large segment a to be used as a unit.
2. Construct $q = \frac{1}{3}a$ and $r = \frac{1}{6}a$.
3. Find the binary expansion of q and of r to 6 places.
4. Construct q+r and find its binary expansion.

Discussion

Let us see how we could have used the binary expansions of q and r to predict the expansion of q+r.

Since

$$0.010101a \leq q < 0.010110a \text{ and}$$

$$0.001010a \leq r < 0.001011a$$

it follows that q + r must be at least as large as the smallest possible value of q added to the smallest possible value of r and q + r must be less than a value which exceeds q added to a value which exceeds r.

Expressed symbolically we have

$$\begin{array}{lcl} 0.010101a & q & 0.010110a \\ \underline{0.001010a} & \underline{r} & \underline{0.001011a} \\ \hline 0.011111a & q+r & 0.100001a \end{array}$$

In more detail, the successive dyadic expansions of segments q and r in terms of segment a are given with the corresponding results for q+r.

$$\begin{array}{lclclclclcl} 0.0a & \le & q & < & 0.1a & & 0.0a & \le & r & < & 0.1a \\ 0.01a & \le & q & < & 0.10a & & 0.00a & \le & r & < & 0.01a \\ 0.010a & \le & q & < & 0.011a & & 0.001a & \le & r & < & 0.010a \\ 0.0101a & \le & q & < & 0.0110a & & 0.0010a & \le & r & < & 0.0011a \\ 0.01010a & \le & q & < & 0.01011a & & 0.00101a & \le & r & < & 0.00110a \\ 0.010101a & \le & q & < & 0.010110a & & 0.001010a & \le & r & < & 0.001011a \\ \vdots & & & & \vdots & & \vdots & & & & \vdots \\ 0.0101010\ldots10a & & q & & 0.0101010\ldots11 & & 0.00101010\ldots10a & & r & & 0.001010\ldots11a \end{array}$$

$$\begin{array}{lcll} 0.0a & \le q+r & < & 1.0a \\ 0.01a & \le q+r & < & 0.11a \\ 0.011a & \le q+r & < & 0.101a \\ 0.0111a & \le q+r & < & 0.1001a \\ 0.01111a & \le q+r & < & 0.10001a \\ 0.011111a & \le q+r & < & 0.100001a \\ \vdots & \vdots & & \vdots \\ 0.01111\ldots11a & \le q+r & < & 0.10000\ldots1a \end{array}$$

As we steadily improve the accuracy of the dyadic expansion of q and r , we get better and better estimates on q+r. In our case we see that the dyadic expansion of q+r should be either .01111111... a or .010...0 a and we must agree that these two expansions represent the same number. See the discussion.

We have seen how the addition of segments corresponds to the addition of their corresponding dyadic expansions, once we have chosen a unit: If we start with segments q and r we can find their dyadic expansions, add these dyadic numbers and construct the segment corresponding to the sum. The segment we obtain will be q + r. In this sense we are able to "translate" arithmetic into geometry and vica versa.

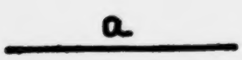
Multiplication of a segment by a real number. We know how to multiply a segment by an integer. For instance 5a = a + a + a + a + a. This multiplication by an integer reduces to repeated addition. We also know how to multiply a segment by 1/2: starting with segment a we simply bisect it to find a segment 1/2 a such that 1/2 a + 1/2 a = a. In this way we know the meaning of .001 a which we obtain by successively bisecting a three times. We then know the meaning of (101.101) x a for instance. It is obtained as 5 a + 1/2 a + 1/8 a. In this way we know how to multiply a segment by finite dyadic expansion. We also know how to multiply a segment by an infinite dyadic expansion: For any dyadic expansion such as r = 1.011010... (which keeps on going) and any segment, c, we can construct r c to any desired degree of accuracy. For instance (taking the above value of r) we know that

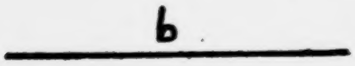
$$1.011010\ c \le rc < 1.011011\ c$$

and so on. If we have a segment b whose dyadic expansion in terms of our unit is r, we can also construct the segment rc geometrically as in the next two experiments.

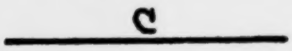
Experiment 23.

1. Choose a unit, a.

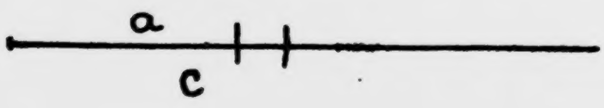




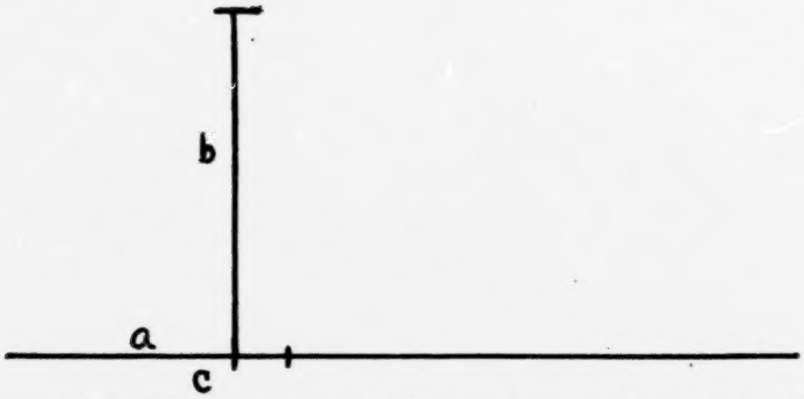
2. Choose a segment b, and find its dyadic expansion in terms of a to five places. Call r this dyadic expansion of b. Thus r = __________ (to five places).
3. Choose a segment c.



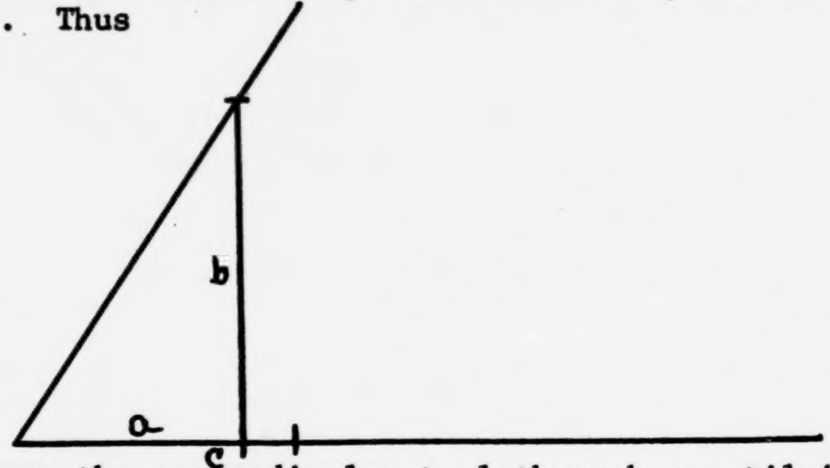
4. On a line, mark off the segment a and the segment c so that they have a common left end point. For instance



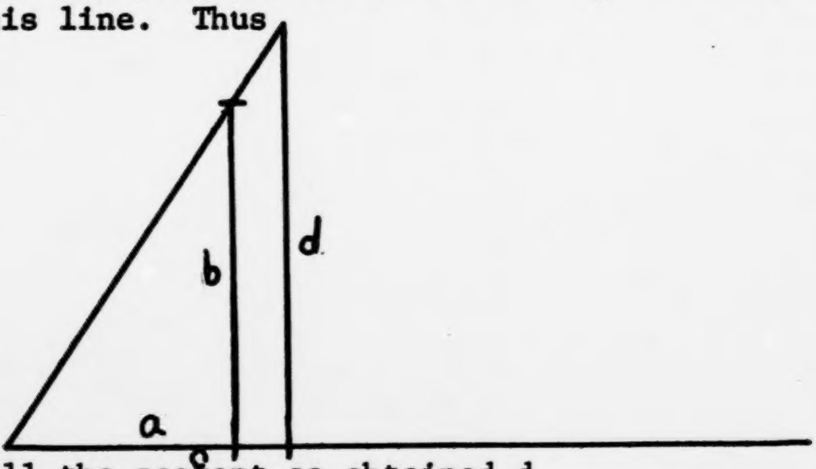
5. Construct a segment equal to b, perpendicular to the line 1, through the end point of a. Thus



6. Draw the line through the other end point of a and of b. Thus



7. Draw the perpendicular to 1 through c until it meets this line. Thus



Call the segment so obtained d.

8. Find the dyadic expansions of c and d in terms of a. Compare r x (the dyadic expansion of c) with the dyadic expansion of d.

Discussion. This way of multiplying is the way used frequently by the Arabs. It has some advantages over direct computation with all the binary expansions, at least in those cases where the binary expansions are rather complicated. We illustrate this in the next experiment.

Experiment 24. Purpose - - to multiply $\sqrt{3} \times \sqrt{5}$.

1. In terms of the unit a construct the segments $b = \sqrt{3}a$ and $c = \sqrt{5}a$ by drawing the appropriate right triangles.

2. Using the procedure of experiment 23 construct $d = \sqrt{3} \times c$.

3. Find the dyadic expansion of d (to five places) and call it s. Compute s^2.

4. Construct $\sqrt{15} \cdot a$ directly via right triangles:

5. Compare $\sqrt{15}$ a with d.

We can use the geometric construction to illustrate some of the laws of multiplication:

Experiment 25. (The distributive law)

1. Choose a unit segment a and segment $b = ra$.

2. Choose segments c and d.

3. Construct $c + d = e$.

4. Find rc and rd and re by the method of Experiment 23.

5. Construct rc + rd and compare it with re.

It is perhaps worthwhile now to pause to list some of the properties and operations we have been studying of lengths and numbers.

A length is not a number. Nevertheless we can add two lengths to get a third and both the associative and commutative laws hold for this addition. We can multiply a length by a real number to get another length. The distributive laws hold for this multiplication. Of course we can also add and multiply numbers to get other numbers. The various laws are listed in the text.

We have also seen how to assign a number to every length (and a length to every number) once a unit has been chosen. If we change the unit, the rule assigning numbers to lengths will change. Let us illustrate how this change works in a simple case. Suppose we start with a as a unit and b = 2a. Thus the number we assign to b (in terms of the unit a) is 2. Suppose we decide to replace a by a' = 1/3 a. Then b = 2a and a = 3a' so that b = 6a'. Thus the number assigned to b in terms of a' is 6. Replacing the unit a by the smaller unit a' = 1/3 a has the effect of multiplying the number assigned to b by 3. We illustrate this in the next experiment.

Experiment 26. Divide class into six equal groups and call them A, B, C, D, E, and F. Sections A, B, and C will work together in the early stages as will sections D, E, and F.

1. Table A

Draw a unit segment and make two copies of it. Give one copy to B and one to C.

2. Table B

   Construct a segment twice the length of A's segment.

3. Table C

   Construct a segment three times A's segment.

4. Table D

   Draw a unit segment and make two copies of it. Give one to E and one to F.

5. Table E

   Construct a segment twice the length of D's segment.

6. Table F

   Construct a segment three times D's segment.

7. Each person now makes a copy of the segment assigned to their work area.

8. Use the segments (m, n, p, q, r) that you were given earlier and get the dyadic expansion of each of these in terms of your new segment. To conserve time and energy, each person could do a different one.

9. Table A

   Compare you expansion of m to B's expansion of m

   n to B's " n

   p to B's " p

   q to B's " q

   r to B's " r

   Is there any generality developing?

10. Table D

    Do the same as 9 with table E.

    Is there any generality developing?

11. Table A

    Compare your expansion of m to C's expansion of m

    n n

Compare your expansion of p to C's expansion of p

q q

r r

12. Table D

    Do the same as 11 with table F.

    Is there any generality developing?

13. Now compare table B with table C and table E with table F. Can you make any generalization?

14. Compare table A with D

    B with E

    C with F

    Can you generalize?

    If not, try to compare the dyadic expansions of the original segments.

TABLE OF CONTENTS

Chapter II. The One Dimensional Vector Space

2.1 Translations on a line....................3

2.2 Directed segments....................5

2.3 The zero translation....................9

2.4 Addition of vectors....................10

2.5 Laws of addition, multiplication by positive reals....................13

2.6 Properties of the zero vector....................14

2.7 Multiplication by -1....................15

2.8 Negative of a vector....................17

2.9 (-1) x (-1) = 1....................17

2.10 Multiplication by a negative number....................18

2.11 The real number system....................20

2.12 Vector laws....................21

2.13 Coordinates on the line....................22

Chapter II

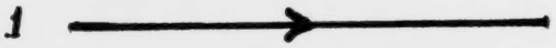
The One Dimensional Vector Space

In Chapter I we studied the number system that enters naturally in the context of measurement. This was the system of "positive real numbers". These numbers do not capture all the meanings we like to associate to a number system. What is missing is a certain symmetry as regards direction. Let us explain what we mean by several examples. When we talk about temperature, we usually express how hot or cold it is by stating the temperature in degrees. We may say that it is 75 degrees or 30 degrees or 10 degrees below zero and so on. The new point here is that we have to talk about "degrees below zero". We never have to talk about a "below zero number of" inches or pounds. If we analyze the situation, we see that the difference is due, in part, to the fact that our notion of "zero degrees" is quite arbitrary. When we talk about weight, it is quite clear to us that an object cannot weigh less than nothing. As to temperature, we feel that it can get hotter and hotter or colder and colder without end. We thus pick some arbitrary point and say that we will measure temperature in both directions from this point. (Actually a deep law of physics says that it can't keep on getting colder - there is an absolutely coldest point. Let us pretend ignorance of this law, however.) We sometimes write $+75^\circ$ for "seventy-five degrees" and -10° for "ten degrees below zero". Notice that in these expressions we have two symbols in addition to the numeral. We have $^\circ$ which signifies "degrees" and either + or - which tells us whether we are "above zero" or "below zero". In this sense, the + or - are not operation signs, and the more correct usage (as found in more recent textbooks) would be to use different symbols than the symbols used for addition and subtraction. No matter how we write it, the expressions of the form -10 occur in many other places besides temperature. For instance, we make talk of an altitude of 200 ft. meaning 200 feet above sea-level and -200 ft. meaning 200 feet below sea level. To say that my bank balance is +$100 means that the bank owes me 100 dollars, while to say that my account is -$50 means that I owe the bank $50. Notice that we can also operate with such expression. To take the last example, let us count a deposit of $20 as a deposit of +$20, while we write a withdrawal of $30 as a "deposit" of -$30. Then starting with $100 in the bank and withdrawing 30 (so we "deposit" -$30) leaves us with $70 in the bank. We can write this as 100 - 30 = 70 or 100 + (-30) = 70. Similarly, starting with 10 dollars in the bank and withdrawing 30 leaves us owing the bank $20, or 10 - 30 = -20.

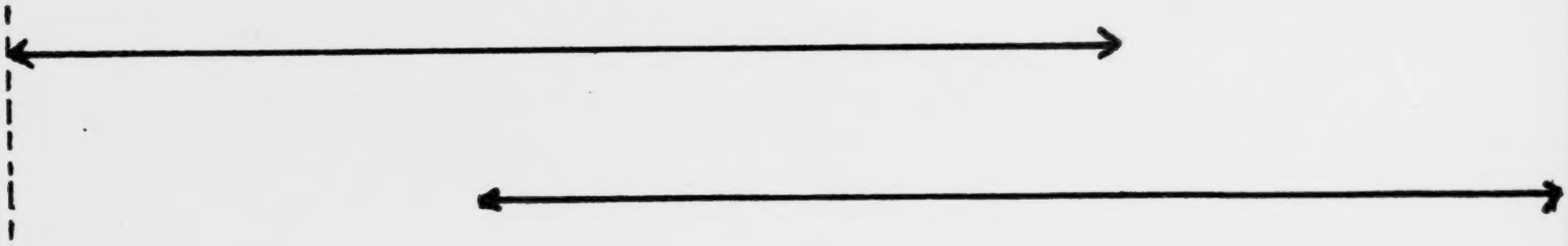
In this chapter we show how "signed real numbers" enter naturally into geometry and study these numbers in the geometrical context. As before, we shall pick a specific geometrical model - this time the study of translations of the line.

2-1. Translations on a Line

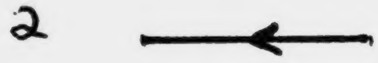
The object we wish to study are sliding motions of a line. That is, we are given a line and can slide it along itself (without changing lengths). We can visualize these motions as on a slide rule, for example. We can slide the inside of the rule in either direction by any amount. (Let us imagine the slide extending indefinitely in both directions.) The things we wish to study are the motions themselves. The first important property about these sliding motions, or as they are called, translations are that we can compare two of them to get a third. Slide the rule once, and then again, the net effect, as far as the change of position is concerned, is the same as making a single translation. Thus translating by this amount and direction



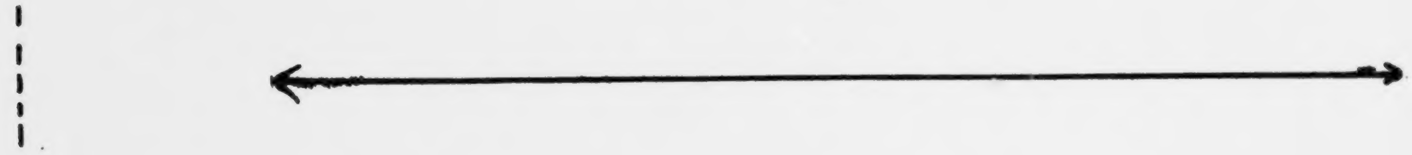
moves the line to the right:



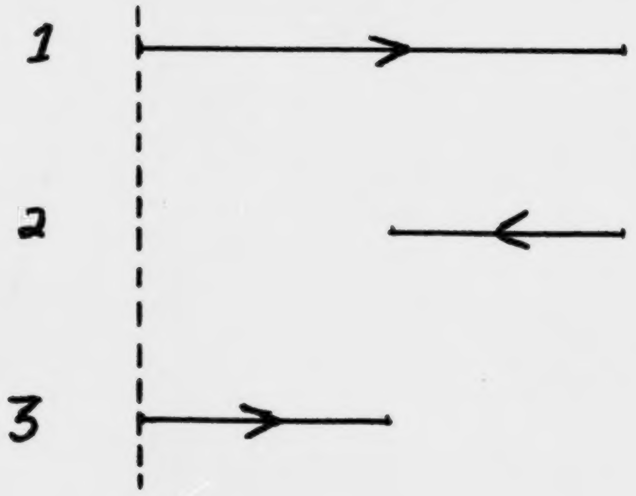
If we then translate by



we move the line to

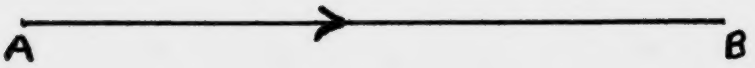


The net effect of the two translations together is the same as a translation by 3



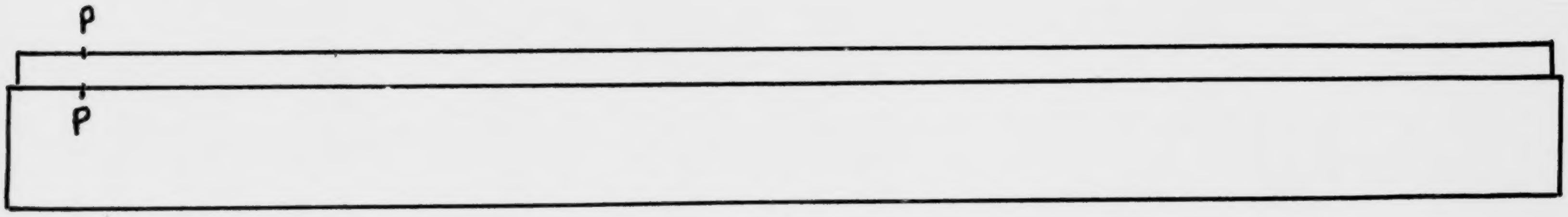
2.2 Directed segments

We want to have some way of labelling and keeping track of our translations. The translations themselves are rather "abstract" objects. They are rules, telling us how to move the line or the slide. To have a more concrete way of dealing with them, we shall proceed as was suggested by the diagrams in 2-1. We draw a separate line and agree that every directed segment on this line is to represent a translation in the following way: A directed segment is just a segment with an arrow drawn on it so that it has a head and a tail:

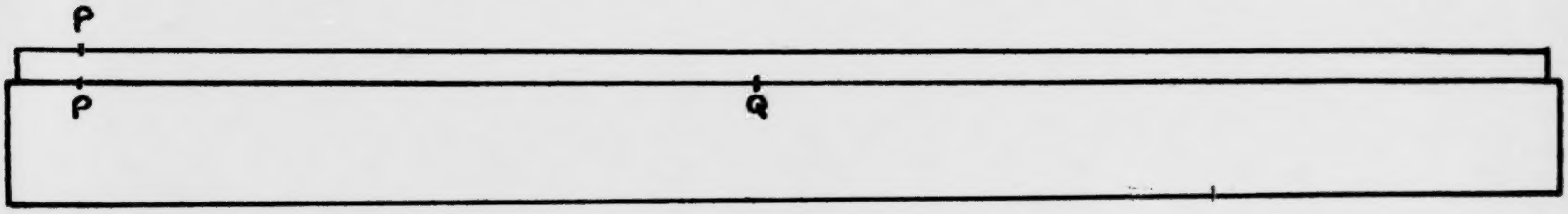


(A is the tail and B is the head)

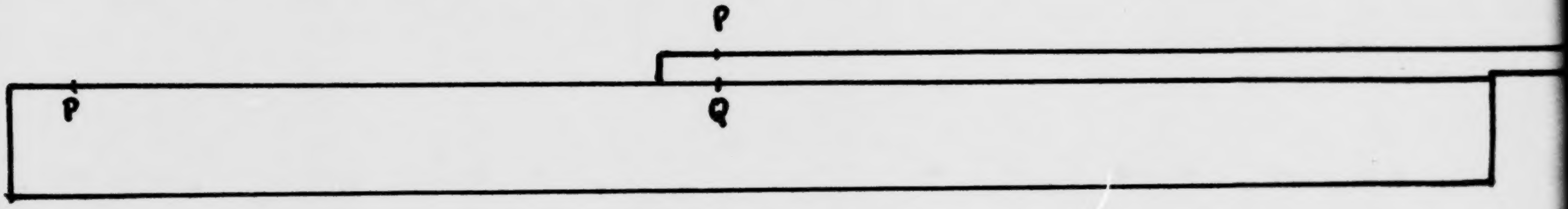
Suppose we start with the directed segment above. Pick any point on the slide rule.



Mark the point, P, both on the slide and on the base. Now reproduce the segment AB on the base putting the tail at P. Call the other end point Q.

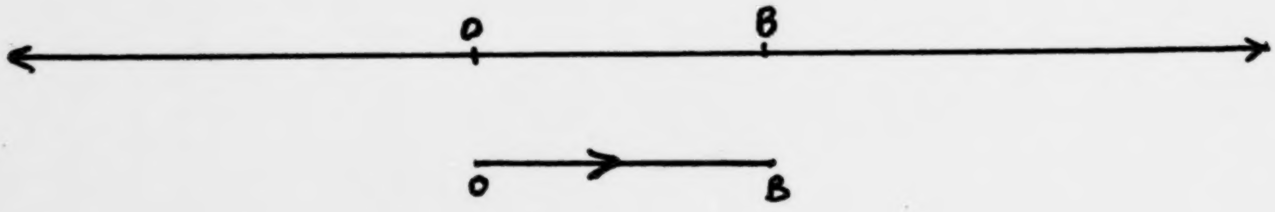


Now move the slide so that the point originally over P now lies over Q.

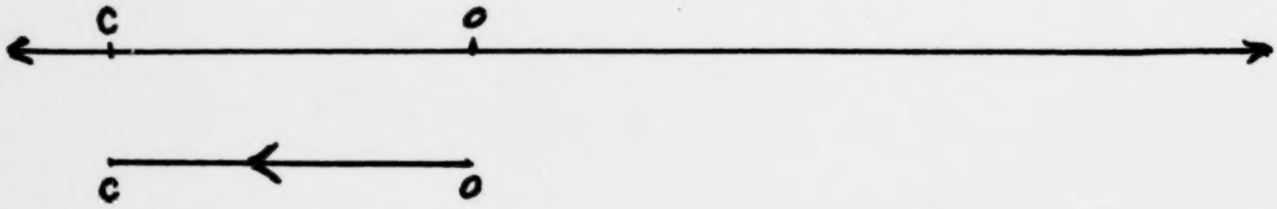


We have thus shown how the segment AB prescribes a motion of the slide. Of course, we must check that the prescription does not depend on which point P we chose. This must, and can, be checked. If we pick some other point as our "start" position, we will find that we will have moved the slide exactly the same way. In fact, it is clear, that the motion of the slide is determined by the length of the segment AB and the direction of the arrow. For this reason we shall be more specific and draw all our segments with a common "tail" point which we shall call '0'. That is, we draw a fixed line and pick a fixed point (or "origin") on the line.

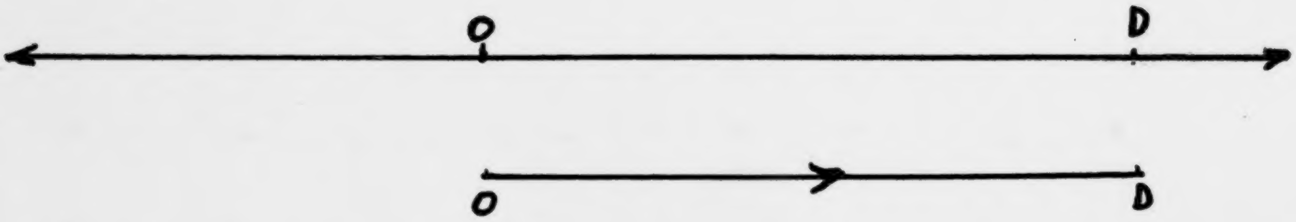
Picking any other point, B, on the line determines a segment OB. Thus,



or



or



are segments with "tail" 'O' and head B, C, or D. Each of the segments OB or OC or OD determines a motion of the slide.

Once we pick the point B or C or D we have determined the segment OB or OC or OD. Each of these segments gives us a rule for moving the slide. In this way, we have associated to each point of the line, a rule for moving the slide. Conversely, start with any motion of the slide. Pick a point P on the slide before the motion. Label Q the point where P ends up after the motion. Then PQ is a segment and we can find a point E such that OE is equal in length to PQ and points in the same direction. Then OE determines the motion we started with.

In this way

> The translations of the slide can be represented as points on the line

2.3 The zero translation

Some special mention should be made of the one special point on our line, the point '0'. What motion does the point '0' determine? A moment's reflection shows that "the motion" corresponding to '0' cannot move the slide at all. In other words, the point '0' corresponds to the "rule of motion" which says "don't move the slide at all." For convenience, we regard this rule as also being a "rule of motion" much the same way as we regard zero as being a number. In fact, we shall call this rule the "zero translation". Its similarity to the number zero will become even more apparent a little later on.

2.4 Addition of vectors

Suppose we are given two translations, v_1 and v_2. We know how to put them together to get a third. The rule is first apply v_1 and then apply v_2. Remember that v_1 and v_2 are directions for moving the slide. We get a new direction which says "first move according to the rule v_1 and then according to the rule v_2". This has the effect of moving the slide and is, in fact, another translation, v_3.

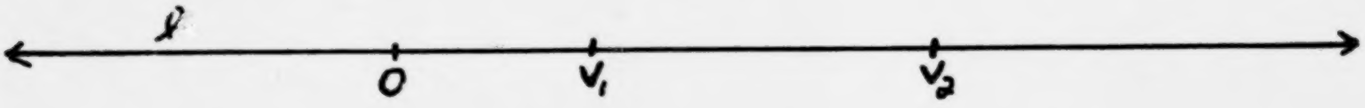
We will denote the operation going from the two translations v_1 and v_2 to the translation v_3 by the overworked symbol, "+". We will thus write:

$$v_3 = v_1 + v_2$$

which says,

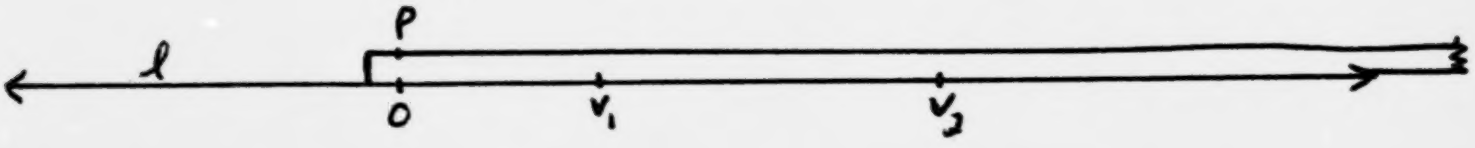
"the translation v_3 is obtained by just moving according to v_1 and then according to v_2".

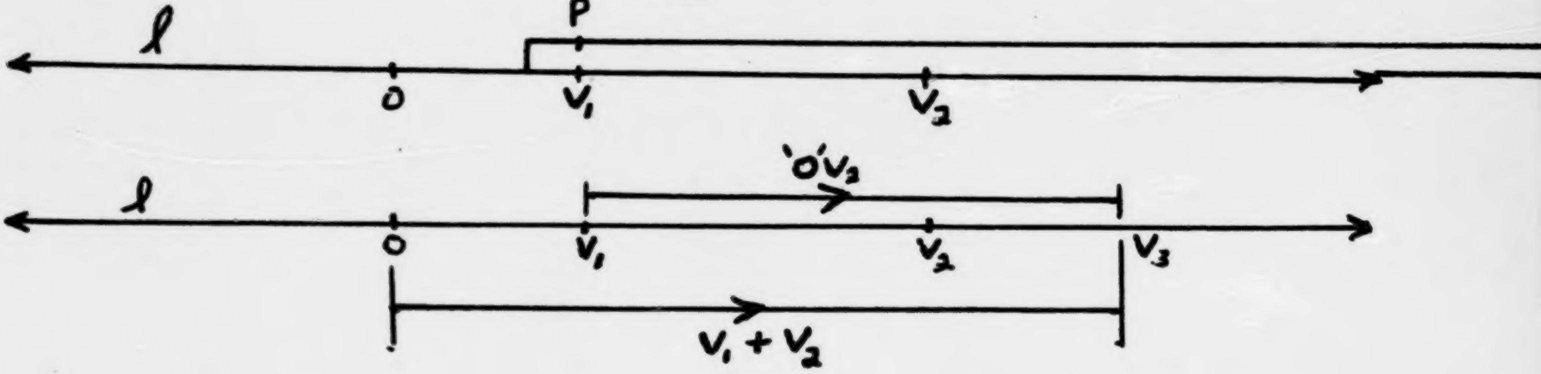
Suppose that we represent the translations v_1 and v_2 as points on our line ℓ Thus, for example, suppose that v_1 and v_2



are given as in the diagram.

How do we find v_3? Imagine our slide is situated with the point P directly over '0'. Then the rule "v_1" says to move P to v_1. Now apply the rule "v_2", picking as our start the point situated over v_1. The rule "v_2" says to draw a segment with tail v_1 equal in length to '0'v_2

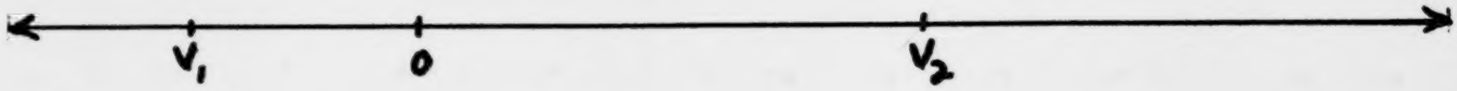


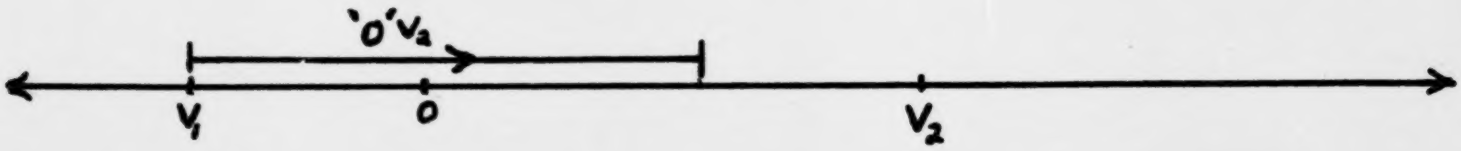

ℓ
P
0
V₁
V₂
ℓ
'O'V₂
0
V₁
V₂
V₃
V₁ + V₂

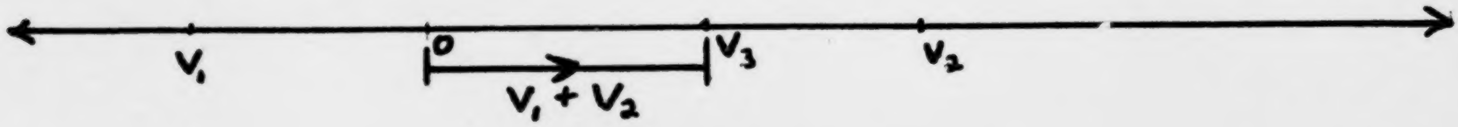
The segment starting at '0' and ending at this new point will correspond to $v_1 + v_2$. To repeat, to find $v_1 + v_2$ on the line we operate as follows:

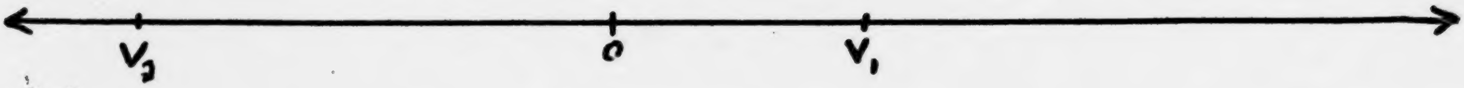
Draw a segment equal in length to '0'v_2 whose tail is v_1. The other end point is $v_1 + v_2$.

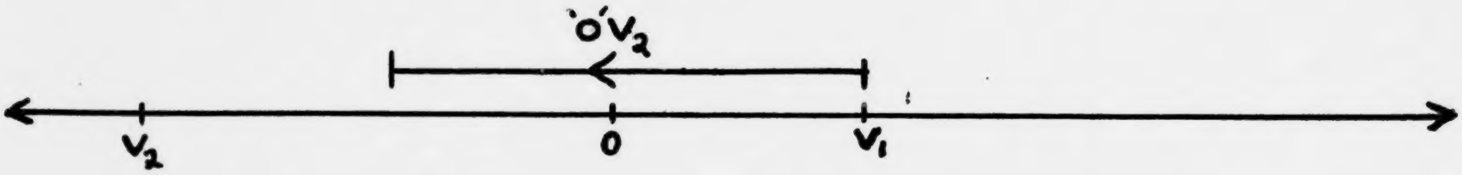
Here are some illustrations:

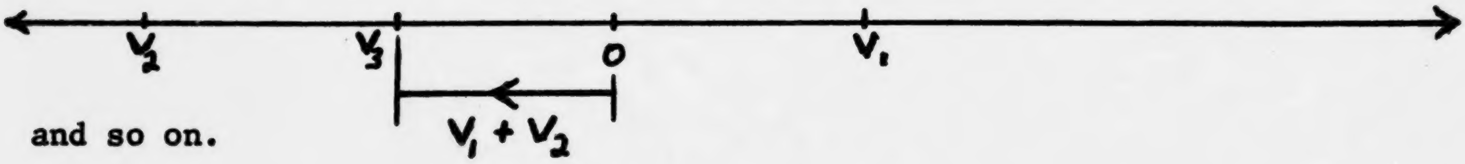





V₁
0
V₃
V₂
V₁ + V₂






V2
V3
0
V1
V1 + V2

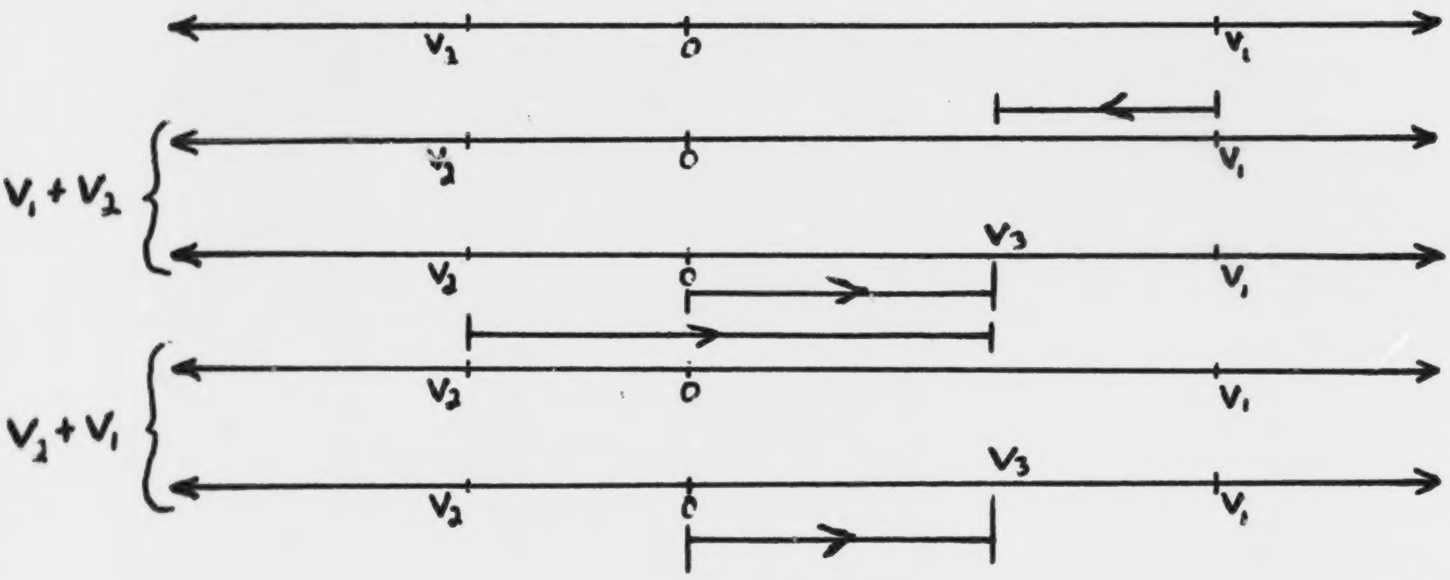
and so on.

2.5 Laws of addition, multiplication by positive reals.

Let us examine some of the properties of this composition. Suppose we start with v_1 and v_2. We can form $v_1 + v_2$ which says "first do v_1 and then do v_2" or we can form $v_2 + v_1$ which says "first do v_2 and then do v_1." Usually, it matters in which order instructions are performed: "first put on your shoes and then your socks" ends you up in a different state of affairs from "first put on your socks and then your shoes." In the present circumstance it doesn't matter.

$$v_1 + v_2 = v_2 + v_1$$

as must be verified experimentally:



Thus the "commutative law holds". Also

$$(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3).$$

In fact the left side says first apply v_1 then v_2 and then v_3 and so does the right. In this case the "associative law" is practically a tautology.

Suppose we start with a translation v . We can form $v_1 + v_1 + v_1$

and we will naturally call this $3v_1$. Similarly, by bisecting the segment '0'v we can find a w such that w + w = v. We will call w = 1/2 v. In this way, we can multiply any translation v by an integer or a dyadic rational, or, in fact, by any positive real number:

> Start with the vector v and the positive real number r. Find the segment whose length is equal to r x (the length of '0'v). Draw the segment of this length with tail '0' and which points in the same direction as '0'v. This will be the translation rv.

2.6 Properties of the zero vector

Let us pay some attention to our special translation '0'. The rule corresponding to '0' is "stay put." If we apply any translation, v, and then stay put this has the same net effect as applying v. Thus

$$v + '0' = v.$$

Thus '0' "acts like zero" as far as "+" is concerned. Since

$$'0' + '0' = '0' \text{ or } 2'0' = '0'$$

and $'0' + '0' + '0' = '0'$ or $3'0' = '0'$ and so on we have n x '0' = '0' for any natural number n. Since '0' + '0' = '0' we know that 1/2 '0' = '0'. Similarly 1/4 '0' = '0' and we make the reasonable conclusion that r'0' = '0' for any positive number r.

We have one further useful convention: The number '0' x any vector = '0'. This coincides with our desire that the distributive law hold:

$$(r + s)\ v = rv + sv$$

if s = '0' and r = 1 we wish to have

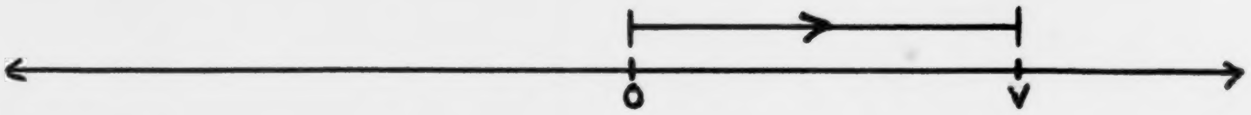
$$(1 +'0')\ v = v + '0'v$$

and we know that v + '0' = v so we get into no trouble by insisting on the rule '0'v = '0'.

We have now reached the point quite close to that of chapter one. We can add two translations and we can multiply any translation by a positive number or by zero, and the usual commutative and associative laws hold for the addition and the various distributive laws hold for the multiplication.

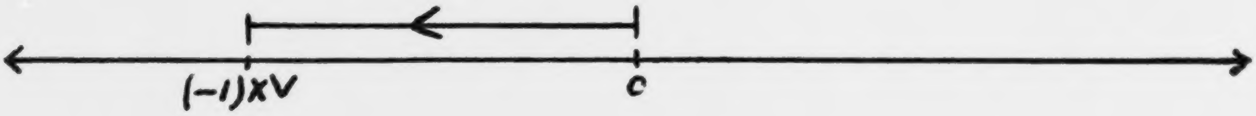
2.7 Multiplication by -1

There is one new operation that we can perform on translations: we can reverse the translation. If a translation carries P into Q we can consider the new translation which carries Q into P. In terms of our representation of translations on our line it says take v and draw the segment of length 'O'v but headed in the opposite direction:



We shall give a name to this operation of reversing v; we shall call it multiplication by -1. Here -1 is just a symbol whose convenience will become more apparent in a little while. Our notation is thus

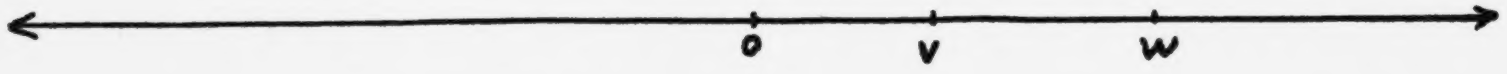
$$(-1) \times v = \text{the ``opposite'' of } v.$$

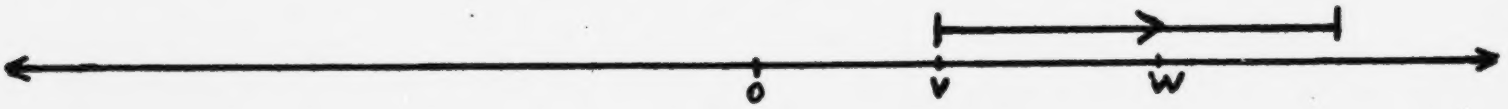


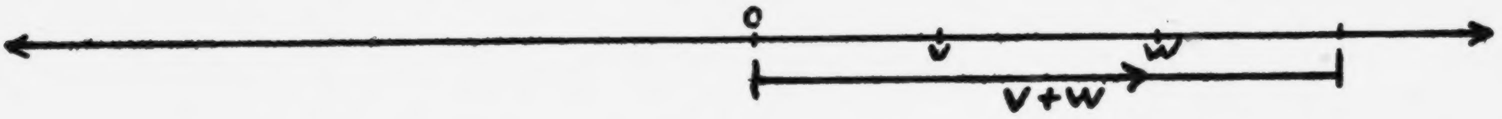
We shall examine some properties of this "reversal" operation. The first one that we take note of is

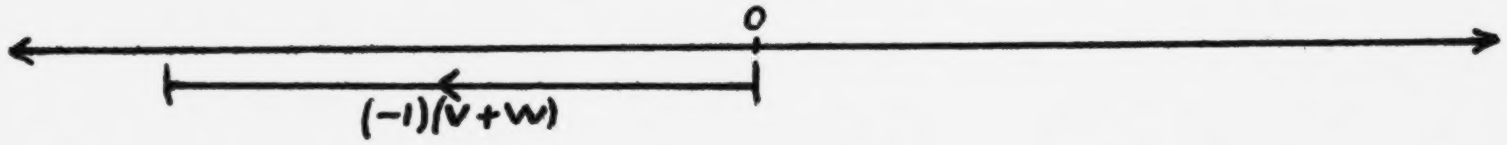
$$(-1)\ (v + w) = (-1) \times v + (-1)\ w.$$

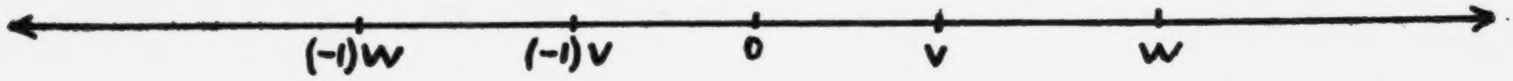
This says that if we first add v and w and then reverse the sum we end up with the same translation as if we had first reversed v and reversed w and then added. We illustrate:

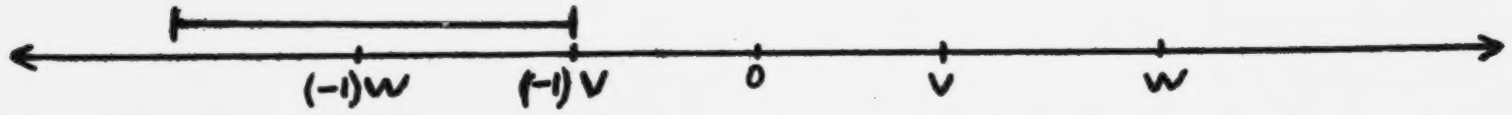


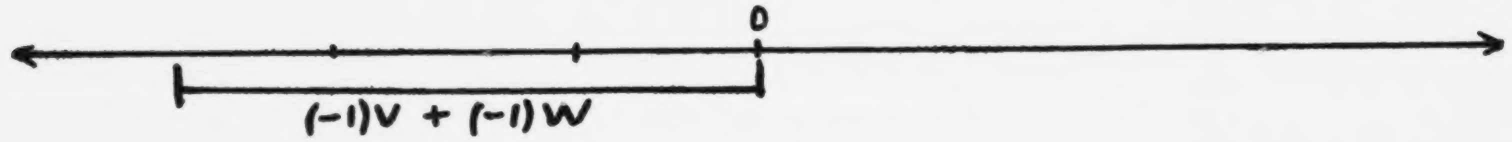












3

2.8 Negative of a vector

The next observation about the reversal operation that we wish to make is perhaps the most obvious one:

$$v + (-1)\ v = \text{'0'}.$$

This says that if we first apply v and then apply the reverse of v we end up back where we started, which is essentially what the reverse of v means.

Notice that this then implies that for any w we can conclude that

$$(w + (-1)\ v) + v = w$$

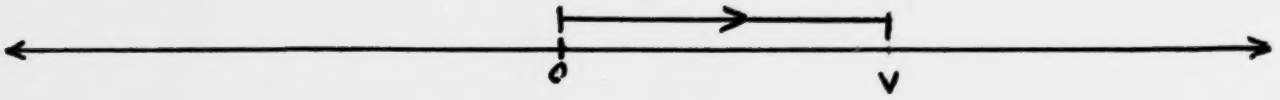
since

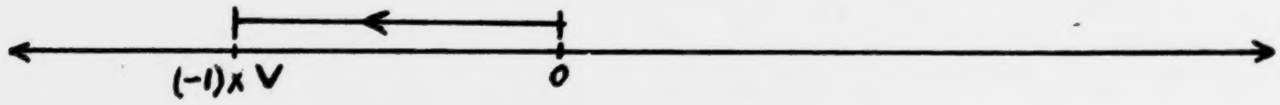
$$\begin{aligned} (w + (-1)v) &= w + ((-1)\ v + v) \\ &= w + \text{'0'} \\ &= w. \end{aligned}$$

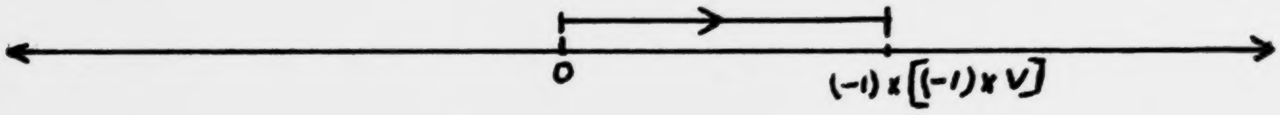
Thus $w + (-1)\ v$ is a translation, which, when v is added to it gives us back w. This is very analogous to the operation of subtraction and we are tempted to write $w - v$ instead of $w + (-1)\ v$. We shall indeed write $w - v$ with the understanding that $w - v$ is a shorthand way of writing $w + (-1)\ v$. For the same reason we shall sometimes write $-v$ as a short way of writing $(-1) \times v$.

2.9 $(-1) \times (-1) = 1$

The next observation about the reversal operation is that reversing a translation twice ends us back with the translation we started with. This is clear both from the definition of the operation and from our geometrical representation.







We write this as

$$(-1) \text{ x } ((-1) \text{ x } v) = v.$$

2.10 Multiplication by a negative number

Suppose we take a v and multiply it by 2. Then reverse the answer. Thus we form (-1) x (2 x v). We know that this is the same as forming 2 x ((-1) x v). We shall introduce some shorthand notation by writing (-2) x v for (-1) x (2v). In other words, we are using the symbol -2 to denote the following operation on v: "double v and reverse the direction." Let us see what the effect of this new notation is. Choose any v. Then

$$(-2) \text{ x } v + v = (-1)\ v + (-1)\ v + v = (-1)\ v$$

or

$$(-2) \text{ x } v + 1 \text{ x } v = (-1) \text{ x } v.$$

If we think of the operation sending v into $(-2) \times v$ as a kind of "multiplication by -2" then this last equation says that for any v if we multiply v by -2 and add the result to what we get by multiplying v by 1 we end up with the same result as multiplying v by -1. This is true for any v. Let us consider the symbols -2, 1, -1 etc. in so far as their effect via multiplication on the v's are concerned. Then we can shorten the previous equation to

$$-2 + 1 = -1.$$

Similarly

$$-5 + 3 = -2$$

$$-4 + 7 = 3$$

$$-2 + 2 = 0$$

in the sense that for any v

$$-5 \times v + 3 \times v = -2 \times v$$

$$-4 \times v + 7 \times v = 3 \times v$$

and

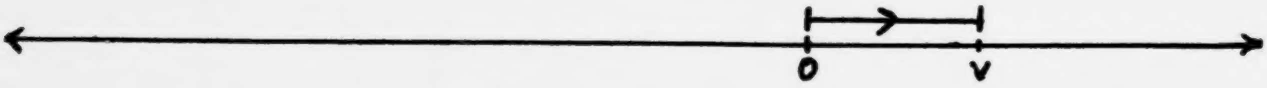
$$-2 \times v + 2 \times v = 0.$$

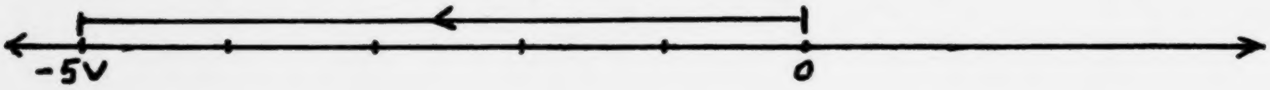
In other words, we are thinking of the symbols 3, 5, -1, -7 and so on as rules for operation v's. As such

$$-5 + 14$$

means the rule which sends any v into $-5v + 14v$ which happens to be the same as 9v. Thus $-5 + 14$ has the same effect as 9 and we write $-5 + 14 = 9$.

Notice that we are already at a double level of abstraction. The v's stand for rules of how to move the slide. We are now studying the symbols -5, 4 etc. which change one v to another. Put another way, the symbol -5 stands for "a change in the rules." We shall not press this point because we prefer to visualize the v's as points on the line. Then -5 is the rule taking any point, v, on the line into -5 v:





The collection of all symbols of the form 1/2, -7, 8, $-\sqrt{3}$ have certain rules of combination. We have studied addition.

2.11

Let us now look at multiplication. If we send v into 2v and then triple the answer we get 6v. In symbols

$$3 \times (2 \times v) = 6 v.$$

We write this as

$$3 \times 2 = 6.$$

If we send v into -2v and then triple the answer we get -6v. In symbols

$$3 \times (-2 \times v) = -6 \times v \qquad \text{or}$$

$$3 \times -2 = -6.$$

If we double v and then multiply by -3 we get -6v, that is

$$-3 \times (2 \times v) = -6v.$$

which we srite as

$$-3 \times 2 = -6.$$

Finally if we multiply v by -2 and then by -3 we have reversed direction twice and multiplied the length of '0'v by 6 and so

$$(-3) \times (-2v) = 6v$$

or

$$-3 \times -2 = 6.$$

We have thus enlarged our "number system" to include all symbols of the form r or -s where r and s are positive (or zero) real numbers. The rules of operation for multiplication are

$$(-r) \times s = r \times (-s) = -(r \times s)$$

and

$$(-r) \times)-s) = rs$$

together with the usual distributive (and commutative and associative laws). The collection of all such members is called the real number system. Thus -5, $\sqrt{14}$, $-\sqrt{3}$, 0, are all real numbers.

2.12 Vector laws

If we are given any translation v and any real number r we can form the new translation rv. We can also add two translations to get a third. Let us collect some of the properties satisfied by these operations. In the following list of properties letters at the end of the alphabet like v, w, z, will stand for translations and letters at the beginning, such as a, b, c, will stand for real numbers:

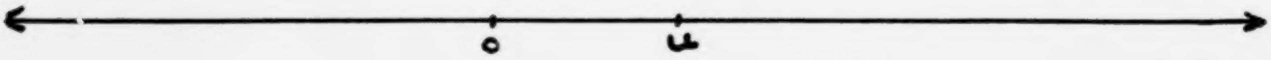
$v + w = w + v$	Commutative Law for addition of vectors
$(v + w) + z = v + (w + z)$	Associative Law for addition of vectors
$'0' + v = v$	Existance of an identity for addition
$a\,(v + w) = av + aw$	Distributive Laws
$(a + b)\,v = av + bv$	
$(a \times b)\,v = a \times (bv)$	

$$1 \times v = v$$

$$-1 \times v + v = 0$$

2.13 Coordinates on the line

Suppose we pick a translation $u \neq 0$ and keep it as our "unit".



We already know from Chapter I that for any v we can find a positive real number r such that the length of '0'v is r x (the length of '0'u). If v and u point in the same direction then $v = r \times u$. If v and u point in opposite directions then $v = -r \times u$. (If v = '0' then v = '0'u.) Thus, once we have chosen our unit u every other v on the line is determined by a real number r. If $v = ru$ and $w = su$ then $v + w = (r + s)\,u$ and no addition of the v's will correspond to addition of real numbers. Similarly for multiplication. In other words:

> Once a unit u has been chosen every v on the line is determined by (and determines) a real number. Addition of the v's corresponds to addition of the real numbers.

We can thus "parametize the line" by the collection of all real numbers.

Callahan/Sternberg/Wei
April 1, 1969

Integrated Mathematics Course
for
Prospective Elementary School Teachers

ANALYTIC GEOMETRY OF THE PLANE

Chapter III 1-94

Laboratory Manual 1-13

TABLE OF CONTENTS

Chapter III. Analytic Geometry of the Plane

3.1 Transformations and symmetries....................1

3.2 Translations and vectors....................7

3.3 Addition of vectors....................11

3.4 The axioms for vectors in the plane....................15

3.5 Lattices in the plane 17

3.6 Coordinatization of the plane

3.7 Affine geometry 21

3.8 Linear transformations....................29

3.9 2 x 2 matrices....................31

3.10 Matrix multiplications....................35

3.11 Matrix addition....................43

3.12 The algebra of matrices....................47

3.13 Multiplicative inverses....................49

**3.14 Solving linear equations 60

**3.15 Eigenvalues and eigenvectors....................64

**3.16 Change of bases....................79

3.17 Conformal transformations and complex numbers....................88

Analytic Geometry of the Plane

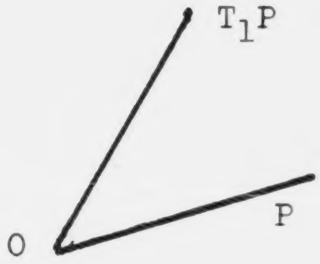
In this chapter we are going to extend the ideas of the last chapter to two dimensions. Our program is to try to express geometrical facts about the plane in algebraic terms. Again, our primary concern will be with building up a certain amount of intuition to make the assertions of linear algebra appear both meaningful and plausible.

The most primitive notion underlying geometry is the idea of transformation. We implicitly think of transformations whenever we are confronted with symmetry. When we see a circle, we notice that rotating the circle about its center leaves it unchanged. When we say that a square exhibits certain symmetries, we mean that there are certain motions of the plane, such as rotating through 90°about the center or flipping over the diagonals which again leave the square unchanged. In short, when we speak of the symmetries of a figure, we mean those transformations we can perform to a plane which do not change the given figure.

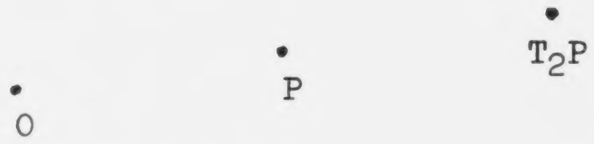
We have been speaking about transformations. What, in fact, do we mean by the word transformation? A transformation is simply a rule, T, which assigns to each point, P, of the plane, another point Tp. (At this juncture we shall not try to define what we mean by the word "plane," and the word "point." For the moment we will get along on the reader's intuitive feelings about these words.) As typical examples of transformations of the plane we mention the following:

1. Let some point O of the plane be fixed. Let T_1 denote the rotation of the plane through a 45 angle about the point O. Thus $T_1O=O$ while if P is some point other than O, the point T_1P is the same distance

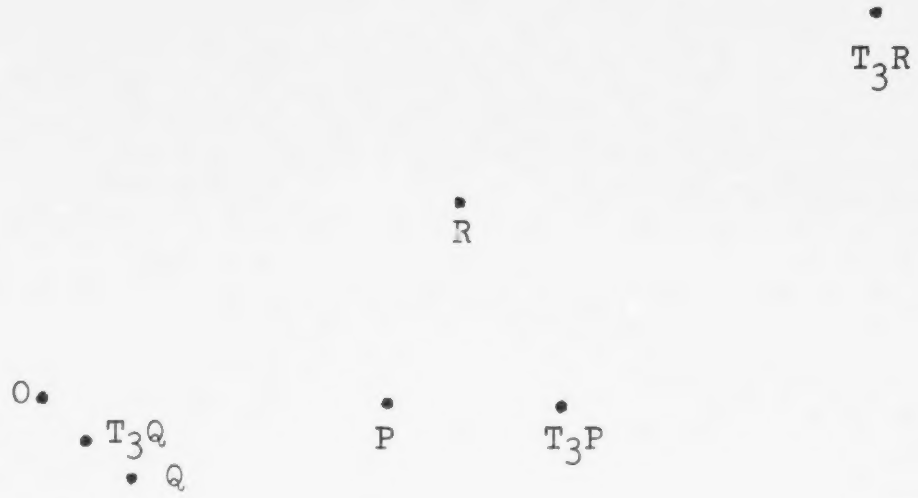
from O as P and the angle PO (T_1P) is 45.



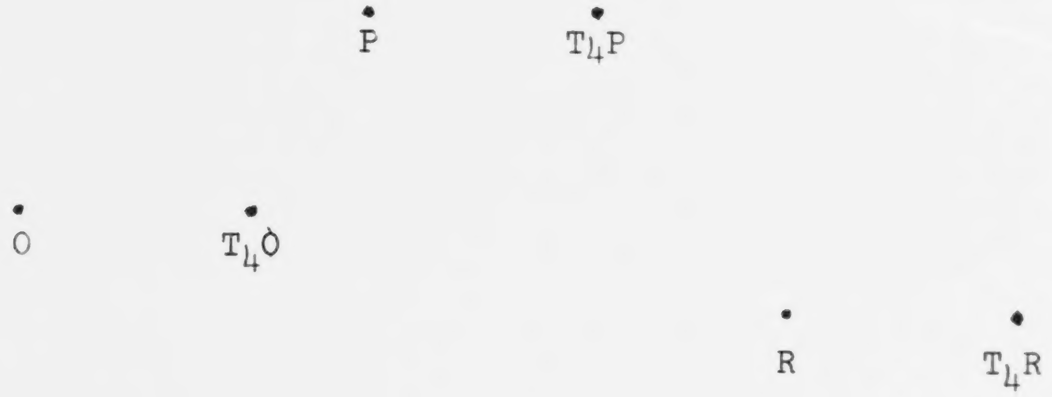
2. Let O be a fixed point of the plane again, and let T_2 assign to O the point O and to each point P different from O the point q which lies on the line from O through P but is twice the distance from O then P is



3. Again, let O be a point, let $T_3O=O$ and let T_3P lie on the line from O through P. But this time let T_3P be the point whose distance to O measured in inches is the square of the distance from O to P. In symbols $\overline{T_3(PO)} = \overline{PO}^2$.

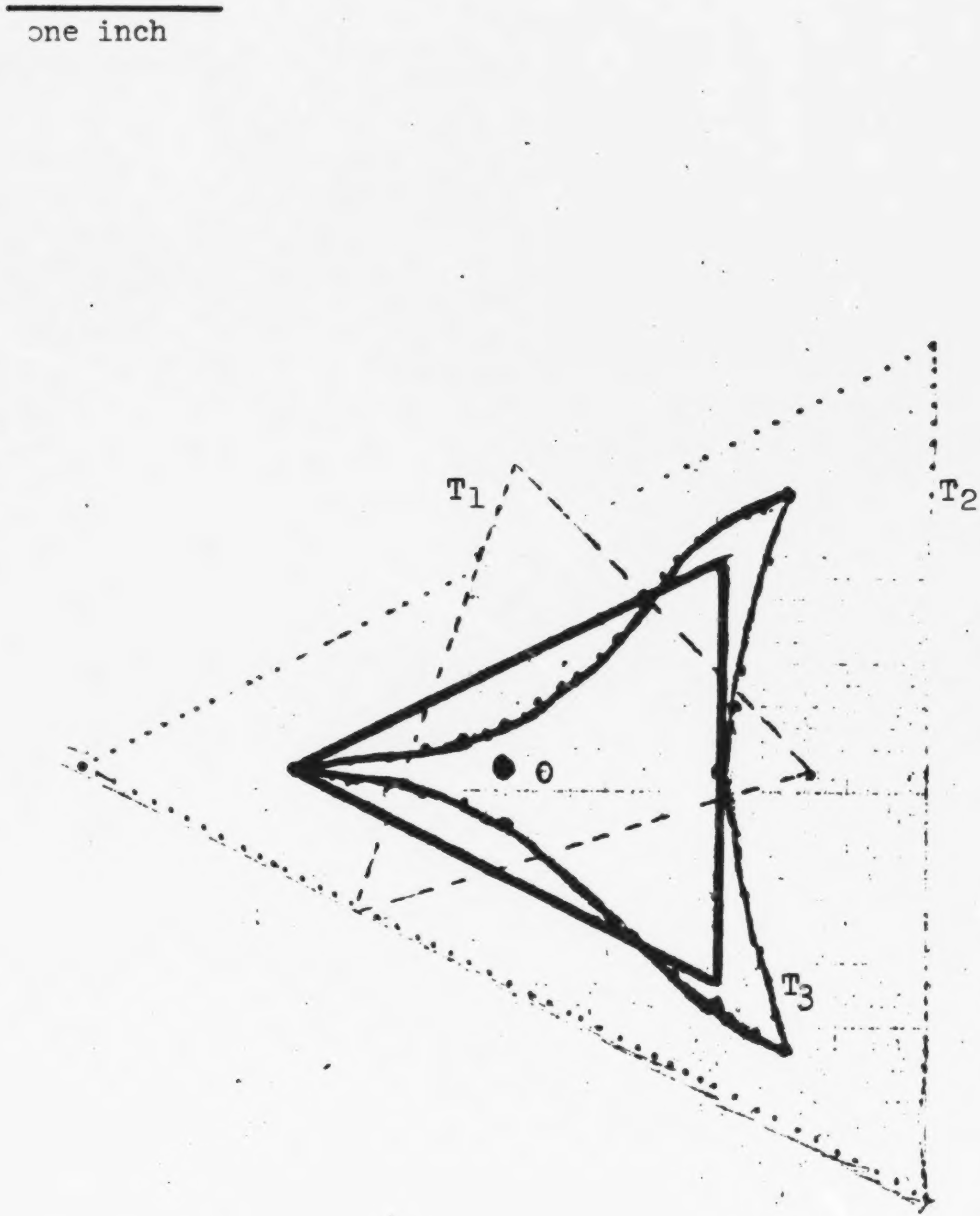


4. As a fourth example, let us assume that our plane comes equipped with directions (NESW) and T_4 consist of moving one inch to the east.



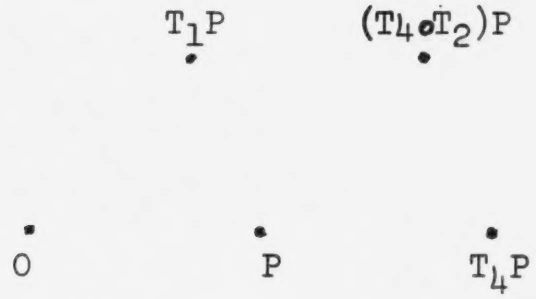
It is easy to imagine more and more complicated transformations of the plane. The reader can easily invent some for himself. It is of interest to see what a transformation does to various figures in the plane. In the following figures we present the result of applying each of the above transformations to a triangle. What is depicted is the result of drawing all the points T_p where the p's are all the points on the triangle.

Thus the dark triangle in the figure is our original triangle. We have drawn the image of this triangle under each of the transformations T_1, T_2 and T_3. Notice that while the image of the triangle under T_1 and T_2 are again triangles, the image under T_3 is not a triangle.



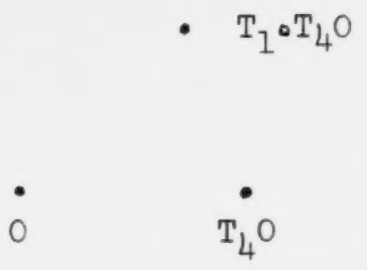
The reader can readily construct or imagine many transformations of his own. He will soon be convinced that one can conceive of some pretty wild and complicated transformations.

What operations can we perform with transformations? The most obvious operation that springs to mind is that of composition. Let S and T be two transformations. We can then consider the composite transformation $T \circ S$ which says first apply the transformation S and then apply the transformation T to what results. Thus



For instance, let us consider the transformation T_4 T_1 where T_4 and T_1 are the transformations given in the previous examples. Then $T_4 \circ T_1$ says first rotate the plane by 45° and then shift to the east by one inch.

Notice that $T_4 \circ T_1$ is not the same as $T_1 \circ T_4$. The transformation $T_1 \circ T_4$ says *first* move one inch to the east and then rotate about the point which now occupies the spot O. To check that these are not the same thing, let us examine where these transformations move the point O.



Thus $(T_4 \circ T_1) O$ is the point lying one inch to the right of O. On the other hand

$$(T_1 \circ T_4)\, O = T_1\, (T_4 O)$$

Now $T_d O$ lies one inch to the east of O. The transformation $T_;$ will rotate this by 45 about O. Thus $T_1 \circ T_4 O$ lies one inch to the northeast of O.

In short, the operation of composition is not commutative in general. The order in which we compare two transformations matters very much in the final outcome.

Although the commutative law fails for general transformations the associative law holds. If R, S, and T are any three transformations then $(T \circ S) \circ R$ and $T \circ (S \circ R)$ represent the same transformation. Indeed suppose that for any point a

$$Ra = b$$

while $\qquad Sb = c.$

Then $\qquad (S\circ R)a = Sb = c$ so

$$T\circ(S\circ R)a = Tc$$

while $\qquad (T\circ S)\circ Ra = (T\circ S)b = Tc$ so that

$$(T\circ S)\ Ra = Tc$$

also. Thus $T\circ(S\circ R)$ and $(T\circ S)\circ R$ have the same effect when applied to any point and are thus more identical.

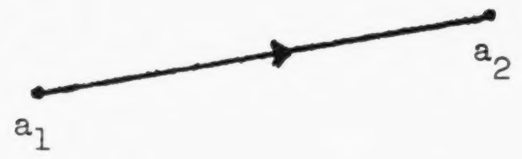
Now the study of all transformations of the plane is a hopelessly complicated mathematical task. In order to be able to make any progress at all, we have to focus attention on a collection of transformations which is manageable from a mathematical point of view. On the other hand, one of our most fundamental feelings about the plane (or about a blank sheet of paper) is that all points are "the same," a feeling that the plane is homogeneous. Thus there should be enough "symmetries of the plane" to carry any point into any other point. Put another way, our collection of transformations should contain enough transformations to move any point of the plane into any other point. We also expect that if two transformations belong to our collection of transformations so should their composition. Otherwise we may have to keep adding transformations to our collection when we compose two transformations. It can be shown that the collection of transformations having all the desired properties and which is simplest in many respects is the collection of all translations.

2. Translations. Intuitively, we can think of a translation of the plane

as a sliding motion of the plane which does not change directions, that is, containing no rotation. We can imagine performing a translation of plane as follows: We can let a sheet of ruled (cross section) paper represent the plane. We place a ruled transparent plastic sheet on the paper so that the rulings match up horizontal lines lying over horizontal lines. In this way we can think of the plastic sheet as simply being another copy of the plane. We now pick up the plastic sheet and place it down in some other position, being sure that the rulings line up once more. In this way, we have "moved" the points of the plane from one position to another, preserving distances and not rotating the plane. Such a motion is called a translation of the plane. Notice that we can carry any point into any other point by a translation: if we have two points a_1 and a_2 on our paper, we can put the plastic sheet down in such a way that the point that used to be over a_1 is now over a_2. (A convenient way to keep track would be to mark the point of the plastic sheet that was originally over a_1 with ink, and now simply place the sheet so that this marked point now lies over a_2.

Notice that not only can we transform any point of the plane into any other point via a translation, there is exactly one translation which will do the job. Given the points a_1 and a_2 there is exactly one translation of the plane carrying a_1 into a_2. (This is one of the ways that the collection consisting of translations alone is a convenient group of transformations to study. If we would allow rotations, for instance, there would be more than one way of transforming the plane which carries a_1 into a_2.)

Since the pair of points a_1 and a_2 determine the translation, we can use them to provide a geometric representation of the translation. In order to indicate which point is moved into which we draw a segment from a_1 to a_2 and put a little arrow on it. Such a segment with an arrow is called a directed segment.



We say that the directed segment a_1a_2 is a representative of the translation taking a_1 to a_2 . Of course, if we started with some other point b_1 we would get a different directed segment, b_1b_2 representing the same translation. Thus while a directed segment determines a unique translation, many different directed segments may determine the same translation. The natural question now arises: when will two directed segments, a_1a_2 and b_1b_2 determine the same translation? The purpose of the first two experiments of this chapter is to convince ourselves of the following fact:

Two directed segments a_1a_2 and b_1b_2 determine the same translation if and only if all of the following three properties hold:

i) the line through a_1 and a_2 is parallel to the line through b_1 and b_2

ii) the distance from a_1 to a_2 is equal to the distance from b_1 to b_2; i.e. $a_1a_2 = b_1b_2$ and

iii) the arrows point in the same direction.

We say that the two directed segments a_1a_2 and b_1b_2 are equivalent if i, ii, and iii are all true. It is easy to check that we have indeed defined an equivalence relation on directed segments. (Notice that we can check properties i, ii, and iii without the use of our plastic sheet, using ruler and compass alone. Thus our equivalence relation makes sense within the confines of Euclidean geometry. We can reformulate the results of our first two experiments as saying that a translation corresponds to an equivalence class of directed segments. An equivalence class of directed segments is called a vector. Thus the word vector is synonymous, for all practical purposes with the word translation.

As usual, special mention must be made of the identity translation. We can consider the transformation of the plane which simply does not move any point as a kind of translation. For reasons of convenience we must consider it in our collection, just as we must count zero as a number. This identity translation carries any point a_1 into a_1 and so does not determine a segment. Nevertheless we can think of the pair a_1a_1 in its own right, and rephrase the previous equivalence relation to read as follows: Two pairs a_1a_2 and b_1b_2 are equivalent if either $a_1 = a_2$ and $b_1 = b_2$ or $a_1 \neq a_2$ and $b_1 \neq b_2$ in which case i), ii), and iii) must hold in order for the pairs to be equivalent. Be it as it may, we have a special kind of vector called the zero vector which corresponds to the identity transformation of the plane.

Addition of Vectors. Let S and T be two translations. We can consider their composite transformations $T \circ S$. The first thing that we notice is that $T \circ S$ is again a translation. The composition of two translations is again a translation. This is our next experimental fact concerning translations. (Notice that there was no reason to expect this in advance. Not all simple looking collections of transformations need be closed under composition. Thus we may consider the following collection of transformations of numbers: We admit any rule which assign to each number n the number $an + b$. The collection of such transformations is closed under composition: sending n into $an + b$ and then sending $an + b$ into $c(an+b) + d$ is the same as sending n into $ac\,n + cb + d$ and is thus another transformation of the same type. But if we consider the collection of all transformations which send n into a number of the form $an^2 + bn + c$ this is not closed under composition. If we consider a second transformation with coefficients e, f, and g then sending n into $e(an^2 + bn + c)^2 + f(an^2 + bn + c) + g$ is not of the same type since it involves an expression raising n to the fourth power.)

The next thing to notice is that if S and T are translations, then

$$T \circ S = S \circ T$$

We verify this fact experimentally in our third experiment. We choose some point O as starting point so that Oa and Ob are directed segments representing S and T. Now choose a as a starting point for a directed segment representing T and choose b as a starting point for a directed segment representing S. It turns out that the end points of these two directed segments coincide. In the language of vectors, let v be the vector standing for S and let w be the vector standing for T. Since $T \circ S$ is again a translation, it corresponds to a vector which we denote by $v + w$ then the equation $S \circ T = T \circ S$ can be written

$$v + w = w + v$$

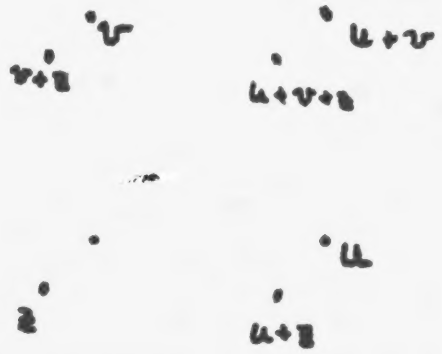
and is known as the commutative law for the addition of vectors. Because of the diagram representing this law, it sometimes is called the parallelogram law.

We now know how to add vectors. We shall be making frequent use of the geometrical representation of the sum of two vectors in what follows. Let us describe the procedure once again. We first choose an arbitrary point O as our origin. Then every vector can be represented by a directed segment whose starting point is O. If Oa represents v and Ob represents w, we can find the directed segment representing $v + w$, and whose starting point is O by constructing the segment parallel and equal to Ob with initial point a (and heading in the same direction as Ob). We could do the construction using ruler and compass, but it is more convenient to use the plastic sheets.

Since the associative law holds for the composition of any three transformations, it certainly holds for the composition of translations. Thus the associative law holds for the addition of vectors:

$$v + (w + z) = (v + w) + z.$$

The following is a diagram illustrating the associative law for the addition of vectors.



In what follows, we shall adopt a convention in drawing the diagrams which illustrate various algebraic laws concerning vectors: We shall fix an origin once and for all in our diagrams. We shall use the same letter to denote the vector and the second end point of the directed segment representing the vector; thus Oa will be the directed segment representing the vector a. The origin will be denoted by 0 since 0 represents the zero vector. In this way, every point in the plane now can also stand for a vector: a point c stands for the directed segment Oc which in turn is a representative for the vector c. The vector c is another name for the translation that sends the point 0 into the point c.

Multiplication of a vector by a number. Let a be a vector. We can form the vectors a + a and a + (a + a) and so on. By the associative law, it doesn't matter how we add various sums of a ,s so that the expression

$$a + a + a$$

is unambiguous, and we denote it, as usual, by 3a. It follows from the associative law that 3a + 2a = 5a, just as we have seen in similar situations in the preceeding two chapters. We can thus talk of the product, na, of a vector by a positive integer, and are able to assert that the distributive law

$$(m+n)a = ma + na$$

holds. By (1/2)a we shall mean the vector which satisfies

$$(1/2)a + (1/2)a = a.$$

If a = 0, then (1/2)a = 0. If a ≠ 0, we can construct (1/2)a by bisecting the segment 0a. If we denote midpoint of the segment by c then indeed c + c = a as can be checked. Similarly we can define the vector ra where r is any positive real number. We could do all of these things by simply mimicking the constructions and definitions of Chapter 2. Actually, we can proceed a little differently. Suppose that a is a non-zero vector. It then determines a line in the plane. Let us consider this line as a one dimensional vector space with origin 0. Then if r is any real number, the vector ra makes sense in terms of the one dimensional vector geometry of Chapter 2. Since ra is a vector lying in a line in our plane, it is a vector in the plane. We can thus consider ra as a vector in the plane. In short, what we are doing, is regarding each line through the origin as a one dimensional vector space.

In any event, we now know how to multiply a vector by any real number. In particular the vector (-1)a has the property that

$$(-1)a + a = 0.$$

We shall therefor denote this vector by -a, just as in Chapter 2.

The one new item that has to be checked is the distributive law for multiplication. If a and b don't lie on the same line, it is no longer a consequence of previous results that

$$r(a + b) = ra + rb.$$

Fortunately this fact is also true and is illustrated by our fifth experiment.

Experiment 6 illustrates a use of the associative law: We are asked to construct $2a + b$ where a and b are such that 2a does not fit on the page. Nevertheless, since $2a + b \;\; 6 = a+(a+b)$ and both a+b and a do lie on the page it turns out that we can find $2a + b$.

The Axioms. We can now state the properties of addition of vectors in the plane in the form of a list of axioms. Except that the symbols now refer to vectors in the plane the axioms are identical with those listed in Chapter 2.

There is a binary operation called addition which assigns to each pair of vector u, and v the vector $u + v$. This binary operation satisfies the

ASSOCIATIVE LAW $\quad u + (v + \;) = (u + v) + w$ for any three vectors u, v, and w

and the

COMMUTATIVE LAW $\quad u + v = v + u$ for any pair of vectors u and v

THE EXISTENCE OF ZERO

there is a vector 0 such that 0 + v = v for any v .

There is also a binary operation called multiplication between real numbers and vectors: given any real number r and any vector v we can form the product rv which is another vector. This multiplication satisfies the

FIRST DISTRIBUTIVE LAW FOR ADDITION (r+s)a = ra + sa for any real numbers r and s and vector v.

and the

SECOND DISTRIBUTIVE LAW FOR ADDITION r(a+b) = ra + rb for any number r and vectors a and b

as well as

ASSOCIATIVE LAW FOR MULTIPLICATION r(sa) = (rs)a for any two numbers r and s and any vector v

and

THE REAL NUMBER ZERO **TIMES** ANY VECTOR IS THE VECTOR ZERO.

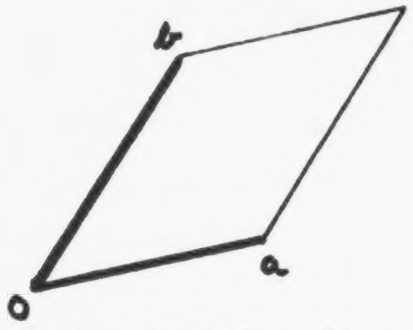
Strictly speaking we should have a separate symbol for the vector 0 and the number 0 . In practice there should never be any confusion whether we are talking about a number or a vector. It is therefor simpler to tolerate some notational ambiguity than to creat a proliferation of symbols.

Linear Independence and Spanning. Let a and b be two non zero vectors that do not determine the same line. Let us start to construct the vectors ma nb for various positive and negative integer values of m and n.

We can construct these points by repeated use of our procedure for adding vector with our plastic sheets. Experiment number eight suggests this method of construction. Actually, since we wish to construct many points of the form ma + nb it is quicker to proceed somewhat differently as described in experiment number nine. There the suggested procedure is as follows: Suppose for instance that we wish to construct all points of the above form where $-5 \leq m \leq 5$ and $-5 \leq n \leq 5$. We first construct the points a, 2a, 3a, 4a 5a, -a, -2a, -3a, -4a -5a using the plastic sheet and do the same for the multiples of b. We then construct the points 5a + b etc. so that we have all points from 5a - 5b to 5a + 5b. Similarly we construct the points a + 5b, 2a + 5b etc. until we have constructed all the points from -5a + 5b to 5a + 5b. We then draw the line from -5a + 5b through -5a, the line from -4a + 5b through -4a and so on until we reach the line from 5a + 5b through 5a. All of these lines are parallel. There are eleven in all. Similarly, we draw the lines from 5a + 5b to 5b, from 5a + 4b to 4b etc., eleven lines all parallel in the direction of a. At the points of intersection we have the various vectors of the desired form.

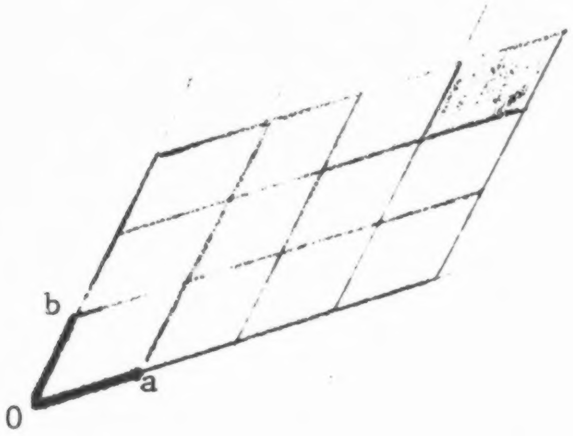
The collection of all the vectors of the form ma + nb where m and n are integers is known as the integral lattice generated by the vectors a and b . It is called a lattice because it looks like lattice work. It is apparent from the picture involved in our construction that we have covered the whole plane with parallelograms whose corners are at the points ma + nb. Every point in the plane lies exactly in one parallelogram, unless it happens to lie on a boundary - a side or a corner of a parallelogram. In this latter case it is ambiguous, which of the two or four parallelograms we should assign to it. This is a problem similar to the problem we encountered in Chapter I.

We shall solve this problem by making a convention analogous to the convention of Chapter I. Let us consider the parallelogram with verticies O, a, b, and a + b. Let us agree that the point O and all the points on the segments join O to a and to b belong to the parallelogram, except for the points a and b themselves. Thus in the diagram, the darkened portion of the boundary belongs to the parallelogram.



It is clear that this convention for one parallelogram then determines what to do for each parallelogram. Each parallelogram now contains, along with its interior, one vertex and two sides. With this convention, each point of the plane now belongs to exactly one parallelogram.

In analogy with our procedure of Chapter I, let us agree to label each parallelogram by the pair of integers describing the vertex it contains. Thus the parallelogram containing the vector $3a + 2b$ will be labelled (3,2).



Now let us construct the vectors (1/2)a and (1/2)b. We can now construct the integral lattice on these vectors or what amounts to the same thing, construct all points in the plane of the form $(m/2)a + (n/2)b$. **See experiment #10** **If we draw the** corresponding parallelograms, we see that we have, in effect, divided each of our previous parallelograms in quarters. Again, with the same convention as before, each point of the plane now belongs to exactly one of these smaller parallelograms.

We can continue the process in complete analogy to the procedure in Chapters I and II. In this way we will assign to each point, c, of the plane, a pair of real numbers (r,s). These numbers are called the coordinates of c relative to the basis consisting of the vectors a and b. We can check that

$$c = ra + sb.$$

This last equation suggests an alternate way of obtaining the numbers r and s from the vector c: Through the point c draw the lines parallel to the lines determined by the vectors a and b. On each of the lines through a and b, mark off the points of intersection with these parallel lines through c. The point of intersection with the line through a is a point lying on the line through a.

We have already remarked that we can consider this line as a one dimensional vector space in its own right. Thus, by the results of Chapter II, we know that this point can be expressed as some multiple of a by a real number. We soon enough discover that this multiple is r. Similarly, the point of intersection lying on the line through b is exactly sb.

We thus have two procedures (which give the same answer) which assign a pair of real numbers to each vector in the plane. Conversely, given the pair of real r and s we can construct the vector ra + sb. Thus, once a choice of a and b is made, every point in the plane corresponds to exactly one pair of real numbers and conversely.

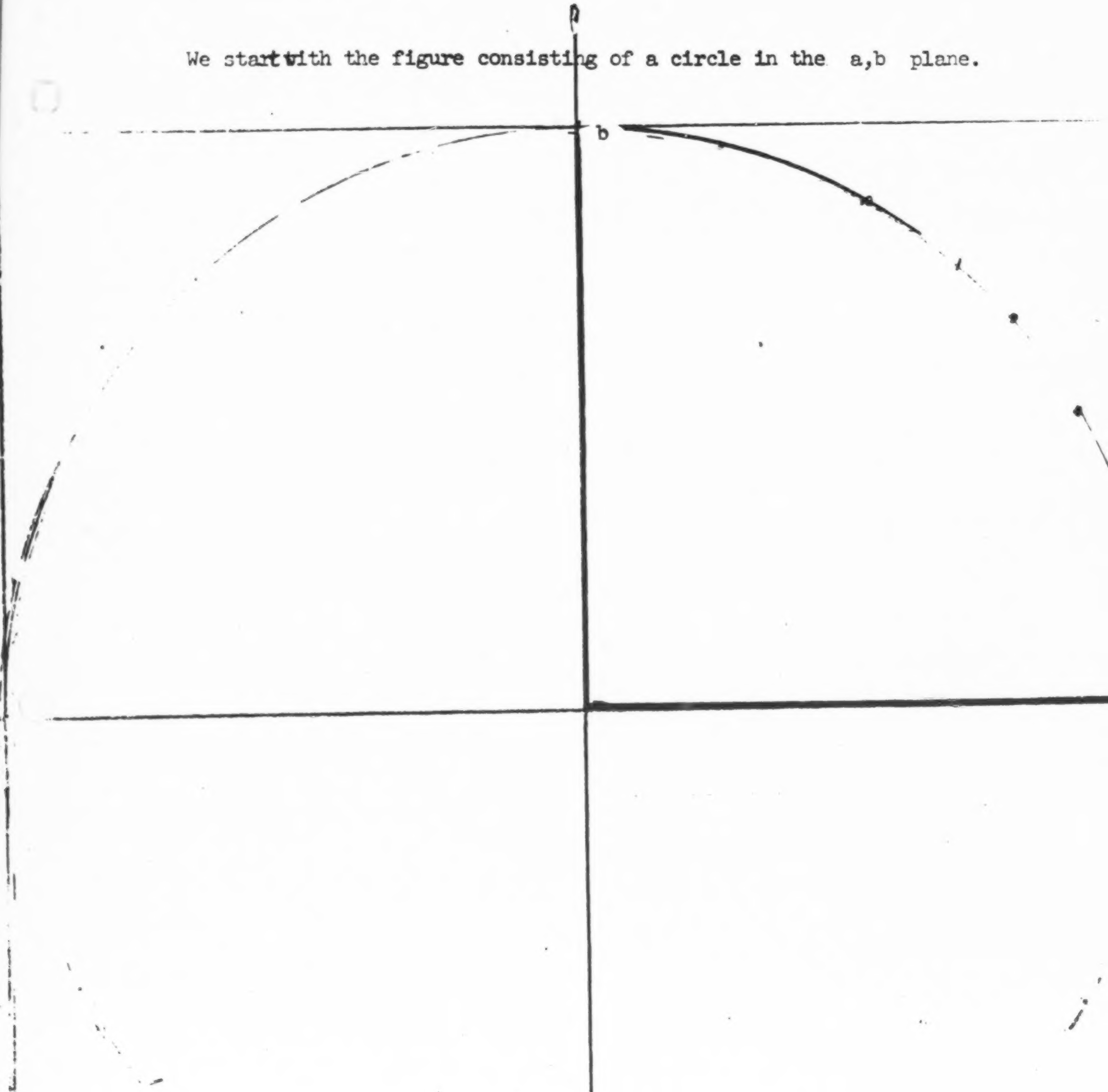
Notice that the whole procedure depends on the assumption that the vectors a and b do not lie on the same line. If a and b do lie on a line, then all the vectors ra + sb will also lie on this line and thus cannot span the whole plane. We include some pictures of what happens when a and b get closer and closer to being collinnear. Notice that so long as they don't actually lie on one line, the vectors ra + sb fill up the whole plane. But notice also that as a and b get closer to being collinear, the actual values needed of r and s needed to describe a given point c in the plane get larger and larger. As we move b more and more into a line with a the lattice points ma + nb seem to fold up like a folding gate.

Experiments 9-12 are concerned with developing some experience with the introduction of coordinates in the plane.

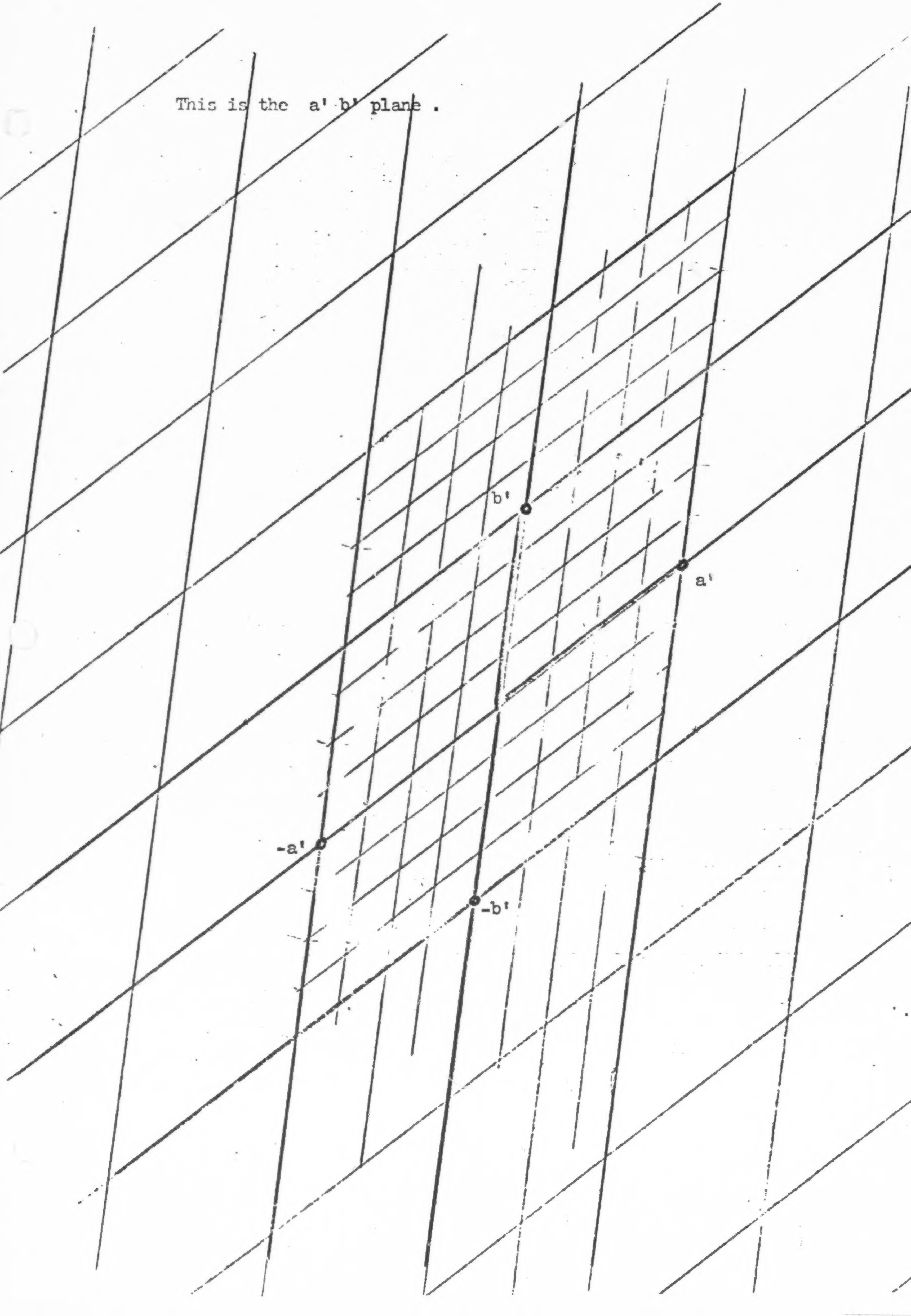
Affine Transformations. We know that the coordinates we introduce in the plane depend on the choice of the basis vectors a and b. In the next sequence of experiments, we wish to investigate the outcome of the following operations: Several different choices of basis vectors are made. For the sake of discussion, suppose that on one plane we choose a pair of basis vectors a and b and a second plane we choose some other basis vectors a' and b' .

Now suppose we draw a figure in the a,b plane. Each point on our figure has certain coordinates. On the a',b' plane, let us draw the points with the corresponding coordinates. In the following sequence of figures, we exhibit the result of performing this with a circle in the a,b plane. The sequence of diagrams shows the result of plotting a few, then several more points in the a',b' plane corresponding to various points on the circle of the a,b,plane. It should be noticed that not only is there a change in the overall scale, there is also a distortion: the image of the circle is definitely not a circle. (It turns out, as we shall see later in the chapter, that the image is an ellipse.)

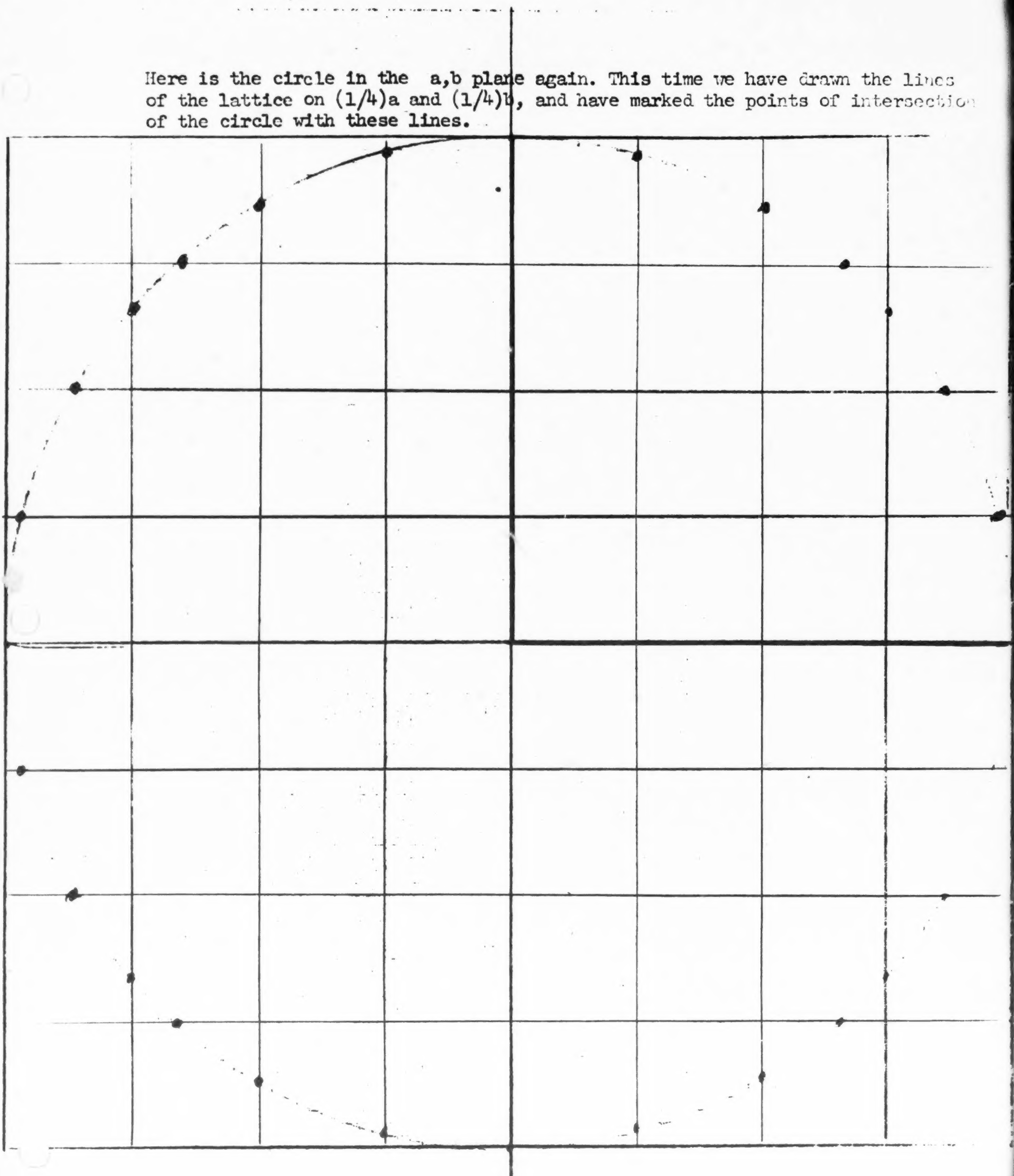
We start with the figure consisting of a circle in the a,b plane.



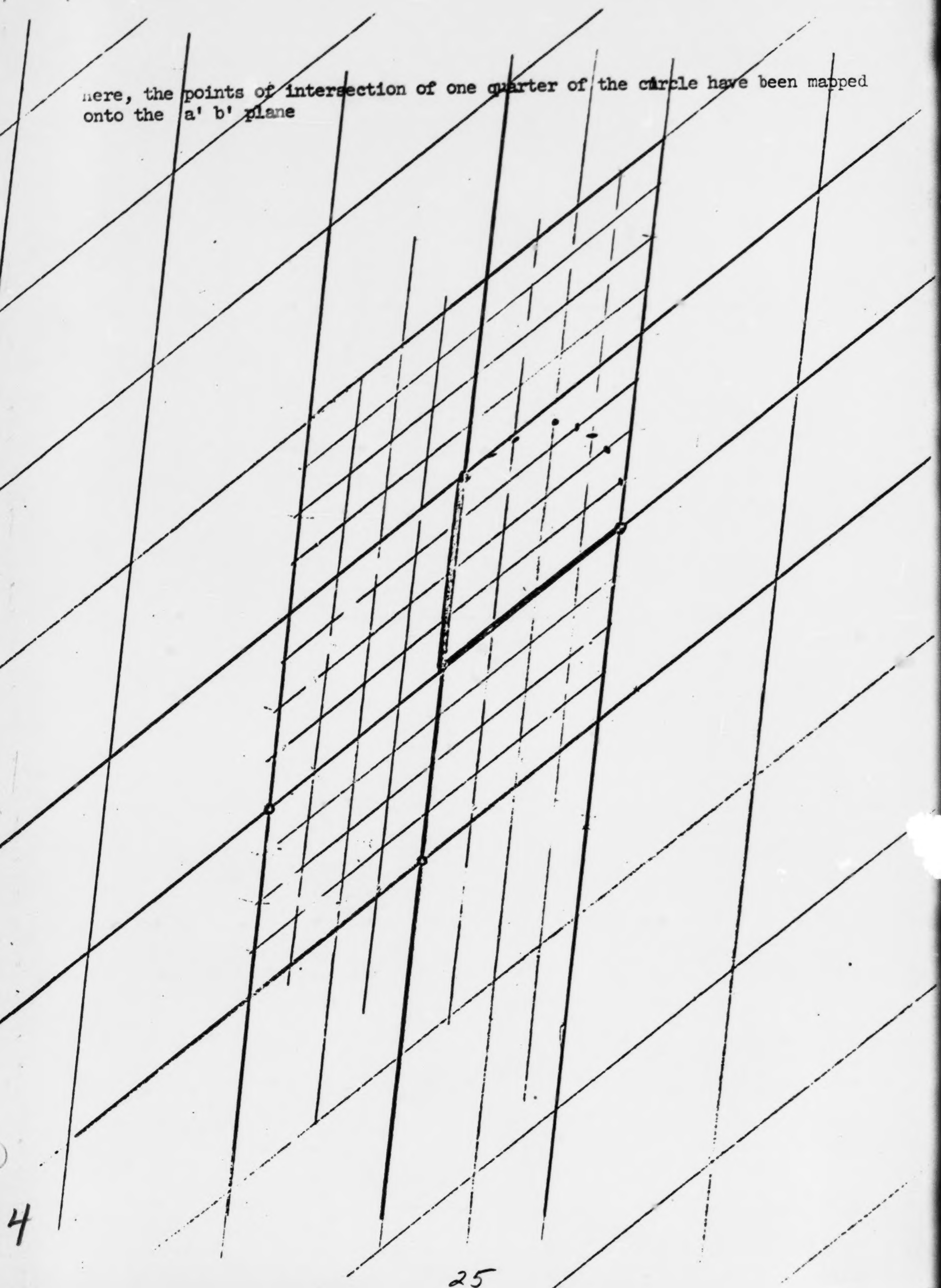
This is the a' b' plane .



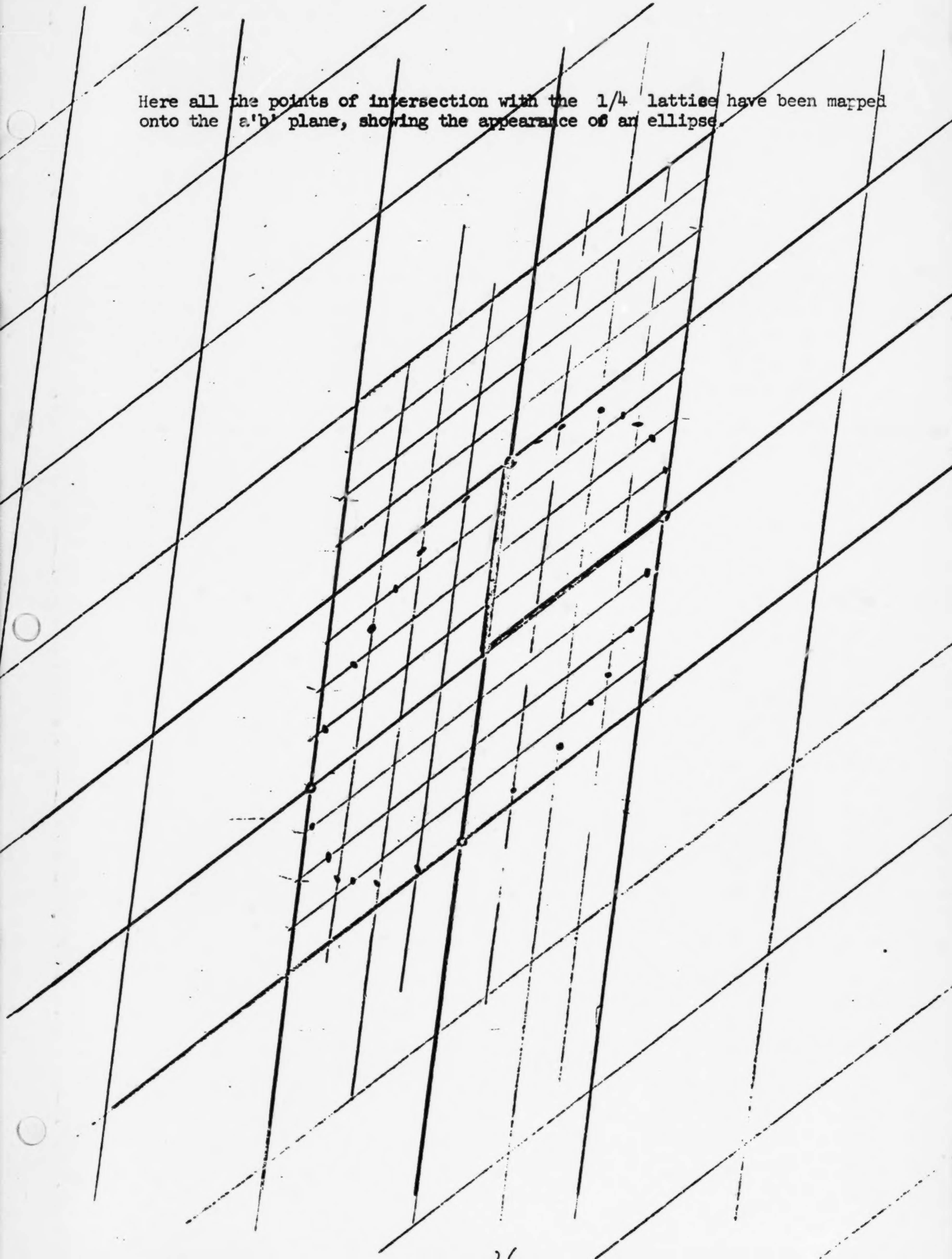
Here is the circle in the a,b plane again. This time we have drawn the lines of the lattice on (1/4)a and (1/4)b, and have marked the points of intersection of the circle with these lines.



here, the points of intersection of one quarter of the circle have been mapped onto the a' b' plane



Here all the points of intersection with the 1/4 lattise have been mapped onto the a'b' plane, showing the appearance of an ellipse.



The purpose of experiments 13 and 14 is to demonstrate the fact that despite the distortion involved in the transferring process from the a,b plane to the a'b' plane, any straight line in the a,b plane is carried over into a straight line in the a'b' plane. (Also, a parallelogram in the a,b plane is carried over into a parallelogram in the a',b' plane.)

We can say that our procedure, which assigns to each point in the a,b plane, a point in the a',b' plane is a transformation from the a,b plane into the a',b' plane. A transformation from one plane to another which carries lines into lines is called an affine transformation. Our experiments show that the transformations we have been studying are affine transformations. A theorem in geometry asserts that the most general affine transformation is obtained by the procedure that we have just described. Let us be more explicit about what this theorem asserts. It says the following: Suppose that T is a one to one transformation of the plane into itself (or on one plane into another). Suppose that T has the property that the image under T of any straight line is again a straight line. Choose three points 0, a and b and let

$$0' = T0, \quad a' = Ta \quad \text{and} \quad b' = Tb$$

If we now apply the above procedure to the vectors 0a and 0b (with choice of origin 0) and to 0' and the vectors 0'a' and 0'b' in the second plane we come up with exactly the transformation T. In this way, our mapping procedure using coordinates constructs the most general affine transformation.

Let us try to see why this theorem should be true. We are starting with a transformation T. All we know about T is that it carries lines into lines and that it doesn't carry two distinct points into the same point. Let f_1 and f_2 be two parallel lines in the a,b plane. Then Tf_1 and Tf_2 must also be parallel lines. They are lines because T carries lines into lines. They are parallel, for if they have a point of intersection, this point would be the image of two distinct points since f_1 and f_2 are parallel. Now let us choose our origin 0 and the two basis vectors 0a and 0b, and define the points 0', a', and b' as we indicated above. Since T carries parallel lines into parallel lines, T will carry the parallelogram spanned by 0, a, b, and a + b into the vertices of a parallelogram. But the fourth vertex of the parallelogram spanned by 0',a' and b' is the point a' + b' (when our origin is 0'). Thus, if we take 0 as origin in the first plane and 0' as the origin in the second, and using these origins, identify points with vectors, we see that

$$T(a+b) = Ta + Tb.$$

But then this same argument shows that T(ma +nb) = mTa + nTb, for all integers m and n. The previous argument may be applied to any pair of vectors, so long as they don't lie on the same line. We can thus apply this result to the vectors $\frac{1}{2^n} a$ and $\frac{1}{2^n} b$. to conclude that

$$T(ra + sb) = ra' + sb'$$

for all dyadic rationals. From this it will follow that the above equation holds for all r and s.

But this last equation says that the transformation is of the type we described above: a point whose coordinates are (r,s) in the first plane is carried over into a point with the same coordinates in the second plane.

Affine geometry is the study of those properties of figures in the plane which are invariant under arbitrary affine transformations. Thus, for example, to say that a quadrilateral is a parallelogram makes sense in affine geometry, since applying an affine transformation to a parallelogram yields another transformation. On the other hand, to say that a quadrilateral is a square makes no sense in affine geometry, because applying an affine transformation we can change a square into an arbitrary parallelogram. Similarly, it makes no sense in affine geometry to say that a figure is a circle, since an affine transformation will, in general covert a circle into an ellipse. (If we consider a circle as a special kind of ellipse, one whose axes are equal, then it makes sense in affine geometry to say that a figure is an ellipse. This is because the most general affine transformation will carry an ellipse into another ellipse . This fact is not obvious and needs to be proved).

Linear Transformations. In order to be able to study the transformations of the last section a little more closely, it is convenient to consider those affine transformations of a plane into itself which keep the origin fixed. Keeping the origin fixed is a minor restriction, because once we are in the same plane, we can always shift the origin back via a translation. An affine transformation of the plane into itself, keeping the origin fixed is called a linear transformation. The rest of this chapter will be devoted to the study of linear transformations of the plane.

Let T be a linear transformation. Let a and b be vectors in the plane. By the results of the last section, we know that

$$T(ra+sb) = rTa + sTb.$$

Thus, if we know the image of a and of b under the linear transformation T, we know the image of any point in the plane. Now the vector Ta lies in our plane, and so has coordinates relative to the basis given by a and b. The values of these coordinates determines Ta and Tb, and thus determine the value of T on any point in the plane. Let us illustrate this by a specific numerical example.

Suppose that $Ta = a + 2b$

and $Tb = -a + b$.

Then

$$T(ra+sb) = rTa + sTb = r(a+2b) + s(-a+b) = (r-s)a + (2r+s)b.$$

Thus, for instance, taking r=1 and s=1.

$$T(a+b) = (1-1)a + (2+1)b = 3b.$$

Similarly, taking $r = -1$ and $s = -2$ we see that

$$\begin{aligned} t(-a \ - 2b) &= (-1)\ Ta + (-2)\ Tb = -(a + 2b) + -2(-a+b) \\ &= (-1+2)\ a + (-2 -a)\ b = a - 4b. \end{aligned}$$

There is a convenient way of writing these relations which will be very useful for many computations later on. We take the coefficients occurring 6n the equation $Ta = a + 2b$ and write them in a column

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

and write the coefficients occurring in the expression for Tb in an adjoining column so as to obtain the expression

$$\begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$$

This square array of numbers is known as the matrix of the linear transformation T.

Now let us take the coordinates of any point c in the plane, for instance, the point $c = 3a + 4b$. The coordinates of c are 3 and 4 We write these coordinates as a column next to the matrix of T

$$\begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

and multiply as follows: to obtain the entry in the first position of the image of c under T, we take the first row of the matrix, multiply the first element in the first row, by the top entry in the column and add this to the second entry in the first row multiplied by the bottom entry in the column representing c :. Thus

$$\begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}\begin{pmatrix}3\\ 4\end{pmatrix} = 1 \times 3 + (-1) \times 4 = -1$$

We obtain the entry in the second position by the same procedure, using the second row of the matrix this time instead of the first. Thus, we get

$$\begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}\begin{pmatrix}3\\ 4\end{pmatrix} = \begin{matrix}1 \times 3 & + & -1 \times 4\\ 2 \times 3 & + & 1 \times 4\end{matrix} = \begin{pmatrix}-1\\ 0\end{pmatrix}$$

or, in short,

$$\begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}\begin{pmatrix}3\\ 4\end{pmatrix} = \begin{pmatrix}-1\\ 10\end{pmatrix}$$

which tells us that the coordinates of the image of c are -1 and 0 if the coordinates of c are 3 and 4. Similarly, if c = -a + 2b

$$\begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}\begin{pmatrix}-1\\ -2\end{pmatrix} = \begin{pmatrix}1 \times (-1) + & -1 \times -2\\ 2 \times (-1) + & 1 \times -2\end{pmatrix} = \begin{pmatrix}1\\ -4\end{pmatrix}$$

For the general point c = ra + sb the computations read

$$\begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}\begin{pmatrix}r\\ s\end{pmatrix} = \begin{pmatrix}1 \times r & + & -1 \times s\\ 2 \times r & + & 1 \times s\end{pmatrix} = \begin{pmatrix}r-s\\ 2r+s\end{pmatrix}$$

In this way, we see that the matrix of a linear transformation gives us the full details on how the transformation actually operates. Let us now formulate the procedure for a general linear transformation. Suppose that T is a linear transformation such that

$$Ta = xa + yb \quad \text{and} \quad Tb = ua + vb.$$

The matrix of this linear transformation T (in terms of the basis vectors a and b) is given by

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}$$

The linear transformation T applied to the vector ra + sb is then computed according to the rule

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} xr + us \\ yr + vs \end{pmatrix}.$$

Exercises. Compute the results of applying the following matrices to the vectors with the given coordinates. In experiments we draw the linear transformation corresponding to some of these matrices.

1. $\begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 1 \end{pmatrix}$

2. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

3. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} -1 \\ 2 \end{pmatrix}$

4. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

5. $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

6. $\begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

7. $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

8. $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

9. $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

10. $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

11. $\begin{pmatrix} 1 & u \\ o & 1 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

12. $\begin{pmatrix} 2 & u \\ 0 & 2 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

13. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

14. $\begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{pmatrix} \times \begin{pmatrix} r \\ s \end{pmatrix}$

Matrix Multiplication, Let S and T be linear transformations. Then T S is again a linear transformations. How can we express the matrix of T S in terms of the matrix of T and the matrix of S. Now this matrix is determined by what the linear transformation $T \circ S$ does to the basis vectors a and b.

Suppose, for example, that S is the linear transformation whose matrix is

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

and that T is the linear transformation whose matrix is

$$\begin{pmatrix} 2 & 3 \\ -2 & 4 \end{pmatrix}.$$

Then

$$S\,a = 1a + 3b\ .$$

We compute the result of applying the linear transformation T to the vector S a as

$$\begin{pmatrix} 2 & 3 \\ -2 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{matrix} 2 \times 1 + 3 \times 3 \\ -2 \times 1 + 4 \times 3 \end{matrix} = \begin{pmatrix} 11 \\ 10 \end{pmatrix}.$$

Thus

$$T\,S\,a = 11a + 10b\ .$$

We obtain the expression for $T\,Sb$ in a similar manner:

$$Sb = 2a + 4b$$

so

$$\begin{pmatrix} 2 & 3 \\ -2 & 4 \end{pmatrix}\begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 16 \\ 12 \end{pmatrix}$$

or

$$T\circ S\,b = 16a + 12\ .$$

We have thus computed

$$T\circ S\,a = 11a + 10b$$

$$T\circ S\,b = 16a + 12b$$

from which we see that the matrix of $T\circ S$ is

$$\begin{pmatrix} 11 & 16 \\ 10 & 12 \end{pmatrix}.$$

We obtained the first column of the matrix

$$\begin{pmatrix} 11 & 16 \\ 10 & 12 \end{pmatrix}$$

by applying the matrix

$$\begin{pmatrix} 2 & 3 \\ -2 & 4 \end{pmatrix}$$

to the vector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ (which is the first column of the matrix of S.

We obtained the second column by applying the matrix $\begin{pmatrix} 2 & 3 \\ -2 & 4 \end{pmatrix}$ to the vector $\begin{pmatrix} 2 \\ 4 \end{pmatrix}$ which is the second column in the matrix $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ of S.

general. Let

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}$$

be the matrix of S and let

$$\begin{pmatrix} x' & u' \\ y' & v' \end{pmatrix}$$

be the matrix of T. The application of the linear transformation S to the vector a (whose coordinates are (1, 0) gives the vector $xa+yb$ whose coordinates are (x,y). Applying the linear transformation T to this vector, we obtain, by our method of computation,

$$\begin{pmatrix} x' & u' \\ y' & v' \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x'x + u'y \\ y'x + v'y \end{pmatrix}$$

in other words,

$$T\,S\,a = (x'x+u'y)a + (y'x+v'y)b.$$

A similar argument allows us to compute the image of b: Since the coordinates of Sb are (u,v) if we apply the operator T to this vector we get

$$\begin{pmatrix} x' & u' \\ y' & v' \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x'u + u'v \\ y'u + v'v \end{pmatrix}$$

We thus see that the matrix of T S is given by

$$\begin{pmatrix} x'x + u'y & x'a + u'v \\ y'x + v'y & y'u + v'v \end{pmatrix}.$$

If we consider the composition of two linear transformations as a sort of multiplication, then we have a formular for the corresponding "product" of two matrices. The product of two matrices is given by the formula

$$\begin{pmatrix} x' & u' \\ y' & v' \end{pmatrix}\begin{pmatrix} x & u \\ y & v \end{pmatrix} = \begin{pmatrix} x'x + u'y & x'a + u'v \\ y'x + v'y & y'u + v'v \end{pmatrix}.$$

To illustrate the meaning of this formula, we shall work a few more numerical examples. The rule for forming the product says to apply the matrix on the left to the first column of the matrix on the right to get the first column of the product matrix; and to apply the matrix on the left to the second column of the matrix on the right to get the second column of the product matrix. Let T have the matrix

$$\begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$$

and let S have the matrix

$$\begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix}.$$

Then, by our previous computation

$$\begin{pmatrix} 1 & -1 \\ & 1 \end{pmatrix} \times \begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + -1 \times 1 & 1 \times (-1) + (-1) \times (-2) \\ 2 \times 1 + 1 \times 1 & 2 \times -1 + (1) \times (-2) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 3 & -4 \end{pmatrix}.$$

Let us now compute the product of these same two matrices in the reverse order. Thus we will be computing the matrix of the linear transformation S T.

$$\begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + (-1) \times 2 & 1 \times (-1) + (-1) \times 1 \\ 1 \times 1 + (-2) \times 2 & 1 \times (-1) + -2 \times 1 \end{pmatrix} = \begin{pmatrix} -1 & -2 \\ -3 & -3 \end{pmatrix}.$$

If we compare this answer with the previous one we see that the matrices we get are unequal. This is just a reflection of the fact the composition of two transformations is not a commutative operation.

In order to gain some experience with the multiplication of matrices it is important to work the **flowing** examples.

1. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$

2. $\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$

3. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 3 & 1 \\ -2 & 1 \end{pmatrix}$

4. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

5. $\begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$

6. $\begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix} \times \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

7. $\begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} \times \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$

8. $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$

9. $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \times \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$

10. $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \times \begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix}$

11. $\begin{pmatrix} x & 0 \\ 0 & q \end{pmatrix} \times \begin{pmatrix} r & t \\ s & w \end{pmatrix}$

12. $\begin{pmatrix} r & t \\ s & w \end{pmatrix} \times \begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix}$

13. $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} x & v \\ y & w \end{pmatrix}$

14. $\begin{pmatrix} x & v \\ y & w \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

15. $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$

16. $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & -5 \\ 0 & 1 \end{pmatrix}$

17. $\begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & u' \\ 0 & 1 \end{pmatrix}$

18. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

19. $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

20. $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \times \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$

The algebra of linear transformations. Until now, all the examples of linear transformations that we have been considering have had the property of being one to one, that is they do not carry two distinct points into one and the same point. In order to proceed further, it is convenient to drop this restriction. This will be necessary if we want to be able to add two linear transformations. In order to be clear on this point, we reformulate our definition of the notion of linear transformation.

A linear transformation is any transformation of the plane with the property that for any pair of vectors a and b and for any pair of real numbers r and s the equation

$$T(ra+sb) = rTa + sTb$$

holds.

As an extreme example of a linear transformation which is not one to one, consider the transformation which sends every vector in the plane into the vector 0. Then this transformation is a linear transformation. In fact, what we have to check is whether or not the above equation holds for all pairs of vectors and all pairs of numbers. It certainly does hold, because if T carries all vectors into 0 then both sides of the above equation are equal to zero no matter what a,b,r, or s actually are.

Let S and T be two linear transformations. We are going to define a new transformation called the sum of these linear transformations, and denoted by $S + T$. To define the transformation $S + T$ we must specify how $S + T$ acts when applied to any vector. We specify it by setting

$$(S + T)a = Sa + Ta .$$

In other words, the transformation $S + T$, when applied to any vector a is simply the sum of the two vectors Sa and Ta. We must check that the transformation $S + T$ is again linear. That is we must check whether

$$(S + T)(ra+sb) = r(S + T)a + s(S + T)b$$

for all vectors a and b and for all numbers r and s. We see that this is indeed true by the following string of equalities:

$$\begin{aligned}(S+T)(ra+sb) &= S(ra+sb) + T(ra+sb) && \text{by the definition of } S + T \\ &= rSa+sSb + rTa +sTb && \text{since } S \text{ and } T \text{ are both linear} \\ &= r(Sa+Ta) + s(Sb+Tb) && \text{by the commutative and distributive laws for vectors} \\ &= r(S+T)a + s(S+T)b && \text{by the definition of } S+T \text{ once again}\end{aligned}$$

Notice the following properties of our notion of addition of linear transformations:

Addition of linear transformations is commutative: $S + T = T + S$

In order to prove this, we must show that both sides of the above equation give the same result when applied to any vector in the plane. But

$$\begin{aligned}(S+T)a &= Sa + Ta \\ &= Ta + Sa && \text{by the commutative law for the addition of \underline{vectors}} \\ &= (T+S)a.\end{aligned}$$

By exactly the same argument, we see that the Associative law for addition of linear transformations: $(S+T) + U = S + (T+U)$ holds.

Let us now consider the result of composing the linear transformation U with the sum $(S+T)$. That is, we wish to examine the linear transformation $U\circ(S+T)$. Applying this to any vector a we see that

$$
\begin{aligned}
U\circ(S+T)\,a &= U(Sa+Ta) \\
&= USa + UTa \quad \text{since } U \text{ is linear} \\
&= (U\circ S + U\circ T)a \ .
\end{aligned}
$$

We thus see that we have

The distributive law $U\circ(S+T) = U\circ S + U\circ T$.

Similarly,

$$
\begin{aligned}
((S+T)\circ U)a = (S+T)(Ua) &= S(Ua) + T(Ua) = (S\circ U)a + (T\circ U)a \\
&= (S\circ U + T\circ U)a \ .
\end{aligned}
$$

In other words, we have the second

Distributive law $(S+T)\circ U = S\circ U + T\circ U$.

Let us call the transformation that takes every vector into the zero vector the zero linear transformation. Thus the transformation 0 is the transformation given by

$$0a = 0$$

for any vector a. It is easy to check that

$$O + T = T \text{ and } O \circ T = O \text{ and } T\ O = O$$

for any linear transformation, T.

Finally, let I be the identity linear transformation. Thus I is the transformation that carries every vector in the plane into itself. Then

$$(I\ T)\mathbf{a} = (T \circ I)a$$

for any vector a, so that we can write

$$I\ T = T \circ I \qquad \text{for any linear transformation } T .$$

Notice that if we regard composition as a sort of multiplication, and define addition the way we have, then the collection of all linear transformations behaves very much like the collection of all numbers. The one striking difference is that the commutative law does not hold in the case of composition of linear transformations. Other than this, our usual axioms for the number system--the associative laws for addition and multiplication, the distributive laws, the commutative law for addition, the existence of an additive identity (a zero) and a multiplicative identity--all of these hold true for the case of the collection of all linear transformations. The only additional law that does not hold for linear transformations is the cancellation law for multiplication. We shall discuss this point in the section after next.

The algebra of matrices. We know that every linear transformation is determined by its matrix once a choice of basis vectors is made. We have already seen how to obtain the matrix of the composite of two linear transformations. What is the formula for the matrix of the sum of two matrices? Suppose that the matrix of S is given by

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$$

and the matrix of T is given by

$$\begin{pmatrix} 4 & 5 \\ 6 & 7 \end{pmatrix}.$$

What is the matrix of S + T ? We must compute (S+T)a and (S+T)b.
Now

$$(S+T)a = Sa + Ta = a + 3b + 4a + 6b = 5a + 9b$$

while

$$(S+T)b = Sb + Tb = 2a + b + 5a + 7b = 7a + 8b \quad .$$

Thus the matrix of S + T is

$$\begin{pmatrix} 5 & 7 \\ 9 & 8 \end{pmatrix}$$

Notice that the rule for obtaining the matrix of S + T from the matrix of S and the matrix of T is very simple: Just add the numbers in the corresponding positions. Thus

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} + \begin{pmatrix} 4 & 5 \\ 6 & 7 \end{pmatrix} = \begin{pmatrix} 1+4 & 2+5 \\ 3+5 & 6+7 \end{pmatrix} = \begin{pmatrix} 5 & 7 \\ 9 & 8 \end{pmatrix}$$

We can now check numerically, in terms of matrices, the various axioms for addition and multiplication. We don't have to do this checking in order to establish that the various laws hold, we know that the operations of matrices reflect the corresponding operations on linear transformations. Thus we are sure that the associative laws for addition and multiplication etc. hold for addition and multiplication of matrices. Nevertheless, let us check some of these laws in order to get some further feeling for addition and multiplication of matrices. For instance, let us check the distributive law by verifying numerically that

$$\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\left[\begin{pmatrix}2 & -1\\1 & 3\end{pmatrix}+\begin{pmatrix}1 & 1\\-1 & 1\end{pmatrix}\right]=\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}2 & -1\\1 & 3\end{pmatrix}+\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}1 & 1\\-1 & 1\end{pmatrix}$$

On the left, the sum inside the parentheses becomes

$$\begin{pmatrix}3 & 0\\0 & 4\end{pmatrix}$$

so that the formula for the product on the left is

$$\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}3 & 0\\0 & 4\end{pmatrix}=\begin{pmatrix}3 & 12\\6 & 4\end{pmatrix}.$$

On the other hand, multiplying the matrices on the right hand side gives

$$\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}2 & -1\\1 & 3\end{pmatrix}=\begin{pmatrix}5 & 8\\5 & 1\end{pmatrix}$$

and

$$\begin{pmatrix}1 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}1 & 1\\-1 & 1\end{pmatrix}=\begin{pmatrix}-2 & 4\\1 & 3\end{pmatrix}$$

Adding these two expressions we get,

$$\begin{pmatrix}5 & 8\\5 & 1\end{pmatrix}+\begin{pmatrix}-2 & 4\\1 & 3\end{pmatrix}=\begin{pmatrix}3 & 12\\6 & 4\end{pmatrix}$$

verifying the distributive law for this special example.

The reader should compute the various matrix products and sums in the following exercises to get some experience with addition and multiplication of matrices.

1. $\left[\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}\right] \times \left[\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}\right]$

2. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \times \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$

3. $\begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix} \times \begin{pmatrix} 4 & 2 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 4 & 2 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix}$

4. $\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \times \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$

5. $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 9 & 8 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 9 & 8 \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

6. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} \times \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Multiplicative inverse. We have already seen that the identity transformation I acts as a unit for multiplication. Of course, the matrix corresponding to the identity transformation is the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The question now arises as to whether we can find a multiplicative inverse for a linear transformation T. That is, we are looking for a linear transformation T^{-1} with the property that

$$T^{-1} T = I.$$

Just as in the case of numbers, we don't expect that every linear transformation will have a multiplicative inverse. For instance, if we consider the zero transformation, 0, then $0\,S = S\,0 = 0$ no matter what the linear transformation S is. Thus 0 cannot have a multiplicative inverse. This is just like the situation with real numbers, the number zero does not have a multiplicative inverse. However, in the case of linear transformations, there will be non-zero linear transformations which will also not have a multiplicative inverse. For example, consider the linear transformation T whose matrix is

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

This is not the zero transformation since its matrix has a one in the upper right hand corner. On the other hand, let us compute the linear transformation T^2. Its matrix is computed by

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

thus $T \circ T = T^2 = 0$. From this it follows that the linear transformation T cannot have a multiplicative inverse. Indeed, suppose that it did and we will derive a contradiction: suppose (contrary to fact) that there is a linear transformation T^{-1} such that

$$T^{-1} \circ T = I .$$

Multiply this equation on the right by T, and using the associative law for multiplication, we see that

$$0 = T^{-1} \circ 0 = T^{-1} \circ T \circ T = I \circ T = T$$

contradicting the fact that T is not 0.

This example also shows that the cancellation law for multiplication does not hold in the case of linear transformations. In fact, we have $T \circ T = T \circ 0 = 0$ but T is not equal to zero.

As another example of a linear transformation which does not have a multiplicative inverse, consider the linear transformation whose matrix is

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} .$$

Let us apply this matrix to the vector $a - b$ whose coordinates are $(1, -1)$. Then we see that

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

In other words, $T(a-b) = 0$. But this implies that T cannot have a multiplicative inverse. In fact, suppose there were a multiplicative inverse T^{-1} to T. Then $T^{-1} \circ T = I$ so that

$$(T^{-1} \circ T)(a-b) = I(a-b) = a-b$$

but

$$T^{-1}\, T\, (a-b) = T^{-1}(T(a-b)) = T^{-1} \cdot 0 = 0 \quad ,$$

which is a contradiction. Thus T does not have a multiplicative inverse.

It is easy to see from a geometric point of view when we would expect a linear transformation T to have a multiplicative inverse and when not. Let us apply T to our basis vectors a and b. Then if Ta and Tb do not lie on the same line through the origin, we can find a linear transformation which takes Ta back into a and Tb back into b. We just use the geometric construction described earlier in the chapter. On the other hand, if Ta and Tb do lie on the same line, then T(ra + sb) = sTa + rTb will also lie on this same line for all values of r and s. Thus T will collapse the whole plane into a line. In such a circumstance we would not be able to find an inverse for the transformation T, since a linear transformation carries a line into a line and not into the whole plane.

We now pose ourselves the following problems: first of all, to determine, in terms of the matrix of a linear transformation, whether or not it has a multiplicative inverse. Secondly, if the transformation does have a multiplicative inverse, to find the matrix of the multiplicative inverse in terms of the matrix of the given linear transformation.

To answer the first question, it turns out that there is a number that we can attach to any matrix. This number has the property that the matrix has an inverse if and only if this number is unequal to zero. Since this number determines whether or not the matrix has a multiplicative inverse, this number is called the determinant of the matrix. As is shown in Experiment 2 this number is closely related to the area of the parallelogram spanned by the vectors Ta and Tb .

We now proceed to give the definition of the determinant of a matrix.

For any matrix

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}$$

form the diagonal products xv and uy and subtract uy from xv, that is form the number

$$xv - uy .$$

This number is called the determinant of the given matrix. For instance, for the matrix

$$\begin{matrix} 1 & 2 \\ 3 & 4 \end{matrix}$$

we form the diagonal products

$$\text{i.e. } 1 \times 4 - 2 \times 3$$

and conclude that the determinant of this matrix is 4 - 6 = -2 . (It turns out that this matrix does indeed have a multiplicative inverse.)

To compute the determinant of the matrix

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

we form the diagonal products

$$1 \times 1 - 1 \times 1$$

and obtain 1 - 1 = 0. Thus the determinant of this matrix is 0. As we have already seen, this matrix does not possess a multiplicative inverse. In experiments--we give a geometric interpretation to the determinant.

Let us now show that if the determinant of a matrix is zero then the matrix cannot have a multiplicative inverse. Thus suppose that the matrix

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}$$

has its determinant zero, i.e. suppose that

$$xv - yu = 0 .$$

We wish to show that this matrix cannot possess a multiplicative inverse. We shall conder several cases when this can occur. Suppose, first of all, that our matrix has the property that $x = 0$ and $y = 0$. That is, suppose our matrix has the form

$$\begin{matrix} 0 & u \\ 0 & v \end{matrix} \ .$$

Then the corresponding linear transformation takes the vector a into zero. By the same argument we gave above for the matrix $\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$, we know that this matrix cannot have a multiplicative inverse.

We may therefor assume that either x or y is unequal to 0. Suppose that x is not zero. Let us apply the corresponding linear transformation, T , to the vector $ua - xb$, which is a non-zero vector because $x \neq 0$. Computing, w obtain

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}\begin{pmatrix} u \\ -x \end{pmatrix} = \begin{pmatrix} xu - ux \\ yu - xv \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Similarly, consider what happens when we apply the transformation T to the vector $va - yb$. We get

$$\begin{pmatrix} x & u \\ y & v \end{pmatrix}\begin{pmatrix} v \\ -y \end{pmatrix} = \begin{pmatrix} xv - uy \\ yv - vy \end{pmatrix} = \begin{matrix} 0 \\ 0 \end{matrix}$$

If $y \neq 0$ the vector $va - yb \neq 0$ and so we have again found a non-zero vector which is sent into zero by T.

We know that this means that T has no multiplicative inverse.

We have thus established in all cases that if the determinant of a matrix vanishes, then the matrix has no multiplicative inverse.

We now must show that if the determinant does not vanish, then the matrix does have a multiplicative inverse. To do this, we shall go one step further and write down a formula for the multiplicative inverse of the matrix in question. Rather than pass immediately to the general formula, let us first explain the formula by numerical examples. The rule for finding the multiplicative inverse is as follows: We start with the given matrix, say

$$\begin{pmatrix} 2 & 3 \\ 2 & 4 \end{pmatrix}$$

and compute its determinant, which in our case is $8 - 6 = 2$. We then take our matrix and switch the entries along the 2-4 diagonal and put a minus sign in front of the entries in the other diagonal, thus obtaining

$$\begin{pmatrix} 4 & -3 \\ -2 & 2 \end{pmatrix}$$

We then divide each number by the determinant, getting

$$\begin{pmatrix} 2 & -3/2 \\ -1 & 1 \end{pmatrix}$$

This last matrix is the multiplicative inverse of the matrix we started with.

To check this, we simply multiply the matrices

$$\begin{pmatrix} 2 & -\frac{3}{2} \\ -1 & 1 \end{pmatrix} \times \begin{pmatrix} 2 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 \times 2 - \frac{3}{2} \times 4 & 2 \times 3 - \frac{3}{2} \times 4 \\ -1 \times 2 + 1 \times 2 & -1 \times 3 + 1 \times 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which is the identity matrix.

Let us work another example. We start with the matrix

$$\begin{matrix} 2 & -1 \\ 1 & 2 \end{matrix}$$

We first compute the determinant, which is $2 \times 2 - 1 \times (-1) = 5$. We then take our matrix

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$$

interchange the elements along the main diagonal and replaces the remaining elements by their negatives, obtaining

$$\begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}.$$

We then divide each term by 5 which is the determinant of the matrix to get

$$\begin{pmatrix} \frac{2}{5} & \frac{1}{5} \\ -\frac{1}{5} & \frac{2}{5} \end{pmatrix}$$

Multiplying out, we check that we have indeed found the multiplicative inverse of the given matrix

$$\begin{pmatrix} \frac{2}{5} & \frac{1}{5} \\ -\frac{1}{5} & \frac{2}{5} \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \times 2 + \frac{1}{5} \times 1 & \frac{2}{5} \times -1 + \frac{1}{5} \times 2 \\ -\frac{1}{5} \times 2 + \frac{2}{5} \times 1 & -\frac{1}{5} \times 2 + \frac{2}{5} \times 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

As another example, let us compute the multiplicative inverse of the matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}$$

Here the determinant is $1 \times 5 - 2 \times 3 = -1$. Interchanging the diagonal terms and replacing the other elements by their negatives yields

$$\begin{pmatrix} 5 & -2 \\ -3 & 1 \end{pmatrix}$$

Dividing each element by -1 finally yields

$$\begin{pmatrix} -5 & 2 \\ 3 & -1 \end{pmatrix}$$

We check that this is indeed the multiplicative inverse by multiplication:

$$\begin{pmatrix} -5 & 2 \\ 3 & -1 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} -5 \times 1 + 2 \times 3 & -5 \times 2 + 2 \times 5 \\ 3 \times 1 + (-1) \times 3 & 3 \times 2 + (-1) \times 5 \end{pmatrix} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

Let us now formulate the rule in general. Let

$$\begin{matrix} x & u \\ y & v \end{matrix}$$

be a matrix whose determinant $d = xv - yu$ is unequal to zero.

Form the matrix

$$\begin{pmatrix} v & -u \\ -y & x \end{pmatrix}$$

and then the matrix

$$\begin{pmatrix} \frac{v}{d} & \frac{-u}{d} \\ \frac{-y}{d} & \frac{x}{d} \end{pmatrix}$$

and we claim that this matrix is the multiplicative inverse of $\begin{pmatrix} x & u \\ y & v \end{pmatrix}$.

We check this by multiplying out:

$$\begin{pmatrix} \frac{v}{d} & -\frac{u}{d} \\ \frac{-y}{d} & \frac{x}{d} \end{pmatrix} \begin{pmatrix} x & u \\ y & v \end{pmatrix} = \begin{pmatrix} \frac{v}{d}x - \frac{u}{d}y & \frac{v}{d}u - \frac{u}{d}v \\ \frac{-y}{d}x + \frac{x}{d}y & \frac{-y}{d}u + \frac{x}{d}v \end{pmatrix} = \begin{pmatrix} \frac{xv-uy}{d} & 0 \\ 0 & \frac{xv-uy}{d} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

In this way we have a formula that provides us the multiplicative inverse of any matrix, provided that we can divide by the determinant, that is, provided that the determinant is not zero. We have thus proved that any matrix with non-zero determinants does indeed have a multiplicative inverse.

To get some feeling for multiplicative inverse, the reader should compute the multiplicative inverse of each of the following matrices, provided that the inverse does exist. If there is no multiplicative inverse then he should indicate this fact. He should check by multiplication that he has indeed found the correct multiplicative inverse.

1. $\begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$

2. $\begin{pmatrix} \text{[illegible]} & \text{[illegible]} \\ 0 & 2 \end{pmatrix}$

3. $\begin{pmatrix} \text{[illegible]} & 0 \\ 0 & 5 \end{pmatrix}$

4. $\begin{pmatrix} x & 0 \\ 0 & v \end{pmatrix}$ where $x \neq 0$ and $v \neq 0$.

5. $\begin{pmatrix} 1 & 0 \\ 0 & \text{[illegible]} \end{pmatrix}$

6. $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$

7. $\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$

8. $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$

9. $\begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$

10. $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

11. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

12. $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$

13. $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$

14. $\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$

15. $\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$

16. $\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$

17. $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$

18. $\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$

Solving linear equations. Computing the multiplicative inverse of a matrix also provides us with a means of solving simultaneous linear equations. We begin our discussion of this point with an illustrative example. Suppose that we start with a linear transformation, say the linear transformation whose matrix is

$$\begin{pmatrix} 2 & 3 \\ 2 & 4 \end{pmatrix}$$

and with a point in the plane, say the point 7a + 11b, whose coordinates are (7,11). Applying our linear transformation to this point gives

$$\begin{pmatrix} 2 & 3 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} 7 \\ 11 \end{pmatrix} = \begin{pmatrix} 2 \times 7 + 3 \times 11 \\ 2 \times 7 + 4 \times 11 \end{pmatrix} = \begin{pmatrix} 47 \\ 58 \end{pmatrix}.$$

In other words,

$$2 \times 7 + 3 \times 11 = 47$$

and

$$2 \times 7 + 4 \times 11 = 58$$

If we apply the inverse matrix to the vector (47, 58) we will, of course, get our original vector (7,11) back again. We have already computed the inverse matrix which is

$$\begin{pmatrix} 2 & -3/2 \\ -1 & 1 \end{pmatrix}$$

Applying it to the vector (47,58), we get

$$\begin{pmatrix} 2 & -3/2 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 47 \\ 58 \end{pmatrix} = \begin{pmatrix} 2 \times 47 - (3/2)\ 58 \\ -1 \times 47 + 1 \times 58 \end{pmatrix} = \begin{pmatrix} 7 \\ 11 \end{pmatrix}$$

does indeed give us back (7,11). Now suppose that someone asked us to find numbers r and s such that

$$2 \times r + 3 \times s = 47$$

and

$$2 \times r + 4 \times s = 58.$$

We could answer by applying the inverse matrix to (47,58) to find that r = 7 and s = 11. For instance, suppose we wish to solve the equations

$$2 \times r + 3 \times s = 5$$
$$2 \times r + 4 \times s = 10.$$

This time we do not know the answer in advance. However, we simply apply the inverse matrix to the vector whose coordinates are (5,10) to obtain,

$$\begin{pmatrix} 2 & -3/2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 10 \end{pmatrix} = \begin{pmatrix} 2 \times 5 - (3/2) \times 10 \\ -1 \times 5 + 1 \times 10 \end{pmatrix} = \begin{pmatrix} -5 \\ 5 \end{pmatrix}$$

The reader can check that (-5,5) is indeed the solution of our pair of equations.

The general procedure is now clear. Suppose we are given the numbers x, y, u, and v, and are also given the numbers e and f. Suppose that we wish to find the unknown numbers r and s satisfying the equations

$$x \times r + u \times s = e$$
and
$$y \times r + v \times s = f$$

If the matrix $\begin{pmatrix} x & u \\ y & v \end{pmatrix}$ has a multiplicative inverse, then we apply this inverse matrix to the vector with coordinates (e,f) we will obtain the vector whose coordinates are (r,s). This then solves our system of linear equations.

If the matrix does not have a multiplicative inverse, the situation is a little more complicated. We know that if the matrix does not have a multiplicative inverse, then the corresponding linear transformation maps the whole plane into a line. If the vector with coordinates (e,f) does not lie on this line, then it can not be obtained by applying the matrix to any vector at all. Thus the system of equations will not have any solution. Thus, for instance, the equations

$$\begin{aligned} 2r + 4s &= 5 \\ 3r + 6s &= 6 \end{aligned}$$

will have no solution at all, because the corresponding transformation maps the whole plane onto the line passing through the vector whose coordinates are (2,3). On the other hand, if the vector (e,f) does lie on the line determined by the matrix, then there will be (many) solutions to the corresponding equation. Thus, in the preceding example, if we had 8 and 12 on the right instead of 5 and 6, so that our equations are

$$\begin{aligned} 2r + 4s &= 8 \\ 3r + 6s &= 12, \end{aligned}$$

We can find many solutions, for instance $r = 4$, $s = 0$, or $r = -1$, $s = 2\frac{1}{2}$ and so on.

This procedure also works in general. Suppose that the matrix $\begin{pmatrix} x & u \\ y & v \end{pmatrix}$ has no multiplicative inverse. Then if all the entries x, y, u and v are all 0, then there is no solution to the equations at all unless e and f are both zero, in which case *any* numbers r and s will do. If at least one of the entries of the matrix is not zero, then the vectors (x, y) and (u, v) must lie on the same line. In this case, the vector (e, f) must lie on the same line, otherwise the equations have no solution. To say that (e, f) lies on this line means that (e, f) is some multiple of (x, y); or, if x and y are both zero, that (e, f) is some multiple of (u, v). Thus, in the example above, (5, 6) does not lie on the same line as (2, 3) and so the equations have no solution. On the other hand, (8, 12) is a multiple of (2, 3), in fact, $8 = 4 \times 2$ and $12 = 4\ 3$ so that the equations do have solutions, for instance, $r = 4$, $s = 0$. Also, (8, 12) is a multiple of (4, 6) since $8 = 2 \times 4$ and $12 = 2 \times 6$, so that $r = 0$ and $s = 2$ is another solution pair for the equations.

The reader should solve the following linear equations (or indicate the lack of solutions or the fact that there is more than one solution). This will provide practice not only in solving linear equations, but also additional practice in evaluating determinants and computing the multiplicative inverses of matrices.

1. $2r = 5$
   $3s = 10$

2. $r + 2s = 5$
   $s = 11$

3. $3r + 2s = 5$
   $2r + 3s = 11$

4. $3r + 2s = 1$
   $2r + 3s = 0$

5. $3r + 2s = 0$
   $3r + 2s = 1$

6. $3r + 2s = 4$
   $3r + 2s = 8$

7. $r + s = 2$
   $2r + 2s = 3$

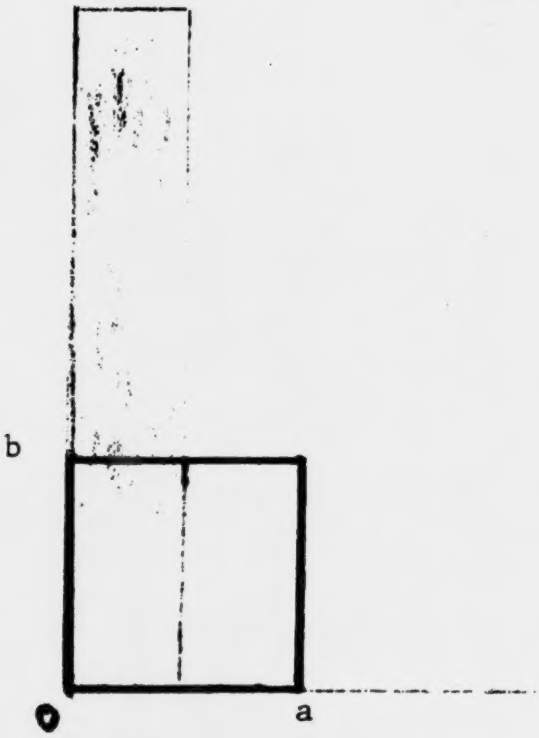
8. $r + s = 10$
   $2r + 2s = 20$

Eigenvalues and eigenvectors. Let us consider a transformation T whose matrix (in terms of our basis vectors a and b) takes the simple form of having non-zero entries only about the main diagonal. Thus suppose, for instance that the matrix of T is

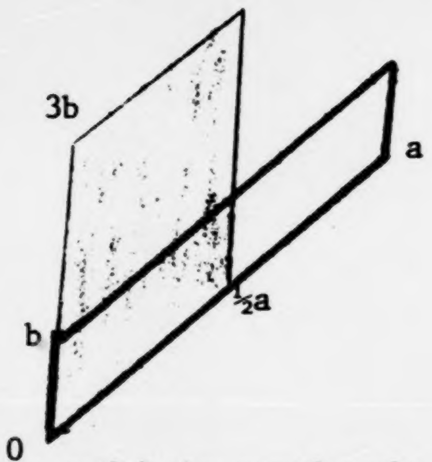
$$\begin{matrix} 3 & 0 \\ 0 & \frac{1}{2} \end{matrix}$$

Such a matrix is called a diagonal matrix. The geometric picture of the action of T is very simple. It stretches everything by a factor of 3 along the a direction and contracts everything by a factor of ½ along the b direction. Thus a diagram of how T acts on a parallelogram whose sides are parallel to the a and b directions is provided below. We give two pictures, corresponding to different choices of basis vectors a and b.

T carries the parallelogram with heavy boundary into the shaded parallelogram



Here is a picture for a different choice of basis vectors, say c and d.



Now the choice of the vectors a and b is completely arbitrary, subject to the condition that they don't lie on the same line. Thus it might be the case that the transformation T carried the vector c into some multiple of itself and carried the vector d into some multiple of itself. If c and d don't lie on the same line, and we had chosen them as our basis vectors, then the matrix of T would be in diagonal form. However, we did not have the good fortune to make this choice of basis vectors. The matrix of T then looks more complicated. Let us illustrate with a specific numerical example. Suppose that the linear transformation T has the property that

$$T(a + b) = 3(a + b)$$

and

$$T(a - 2b) = \tfrac{1}{2}(a - 2b)$$

Thus the transformation T carries the vector a + b in three times itself and carries the vector a - 2b into one half of itself. Let us see what the matrix of the linear transformation T is in terms of our basis vectors a and b. By computing the inverse of the matrix $\begin{matrix} 1 & 1 \\ -1 & -2 \end{matrix}$ we see, or can directly verify, that

$$a = 2/3\,(a+b) + 1/3\,(a-2b)$$

and

$$b = 1/3\,(a+b) - 1/3\,(a-2b)$$

Then

$$\begin{aligned} Ta &= 2/3T\,(a+b) + 1/3T(a-2b) = 2/3 \times 3(a+b) \\ &\quad + 1/3 \times 1/2(a-2b) = (2\ 1/6)a + (1+4/3)b \end{aligned}$$

$$\begin{aligned} Tb &= 1/3T(a+b) - 1/3T(a-2b) = 1/3 \times 3(a+b) \\ &\quad - 1/3 \times 1/2(a-2b) = 5/6a + (1+1/3)b \end{aligned}$$

Thus the matrix of T in terms of the basis a and b is

$$\begin{pmatrix} 2\frac{1}{6} & \frac{5}{6} \\ 1\frac{4}{3} & 1\frac{1}{3} \end{pmatrix}$$

Suppose we had started with this matrix. It certainly looks very complicated. How could we tell that by suitable choice of the vectors c (= a + b) and d (= a - 2b) that the transformation takes the simple form of stretching (or contracting) along the lines determined by these vectors? This is the type of problem which we wish to solve in this section.

Starting out with a linear transformation whose matrix is given to us, we wish to ask the following three questions: First of, are there lines along which the transformation T simply stretches (or contracts) everything? If so, what are the factors of expansion (or contraction)? Thirdly, if our transformation does have this property, what are these lines?

Notice that in formulating the problem we used the word lines instead of the word vectors. The reason is that if T carries the non-zero vector c into some multiple of itself, it will do the same for any other vector lying on the line determined by c.

Observe that not every linear transformation will have the property that it carries some line itself. For instance, if T is rotation through 45° about the origin, then T moves every line and so does not carry any vector into a multiple of itself.

Let T be a linear transformation. We are searching for all possible numbers z and all possible vectors c with the property that T carries the vector c into a multiple of itself by the factor z. In other words, we are looking for numbers z and non-zero vectors c such that

$$Tc = zc$$

It turns out, surprisingly, that we can find out what the possible z's are without knowing the vectors c in advance. Our way of finding the possible z's is to first write the above equation in a slightly different form. Let Z be the linear transformation which multiplies every vector in the plane by the number z. Thus Z is the linear transformation whose matrix is

$$\begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix}$$

We can rewrite the above equation as

$$Tc = Zc$$

or

$$Tc - Zc = 0$$

or, finally

$$(T-Z)c = 0$$

Now we are supposing that c is a non-zero vector. The last equation says that the linear transformation (T-Z) takes this non-zero vector into zero. As we have seen several times already, this means that the linear transformation (T-Z) does not have a multiplicative inverse. This information is sufficient to determine the possible z's. To see how this works, let us examine a numerical example. Suppose that T is a linear transformation whose matrix is

$$\begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix}$$

Then the matrix of T-Z is

$$\begin{pmatrix} 1-z & 2 \\ 5 & 4-z \end{pmatrix}$$

where z is the unknown number we are looking for. To say that T - Z does not have a multiplicative inverse means that the determinant of the last matrix must vanish. We can compute the determinant of this matrix which is

$$(1-z)\ (4-z) \ - \ 2 \times 5 \ = 4-5z + z^2 - 10 \ = z^2 - 5z - 6$$

This means that the number z must make this last expression vanish. Thus z must be a solution of the quadratic equation

$$z^2 - 5z - 6 = 0.$$

Now we can factor this last equation as

$$z^2 - 5z - 6 = (z-6)(z+1).$$

Thus the two possible values of z are z = 6 and z = -1. These two numbers, 6 and -1, are called the eigenvalues of the matrix

$$\begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix}$$

We can indeed check that when we substitute these values for z into T-Z the matrix we get has determinant zero. In fact, taking z = 6 gives the matrix

$$\begin{pmatrix} 1-6 & 2 \\ 5 & 4-6 \end{pmatrix} = \begin{pmatrix} -5 & 2 \\ 5 & -2 \end{pmatrix}$$

whose determinant is 10 - 10 = 0. Also the matrix

$$\begin{pmatrix} 1+1 & 2 \\ 5 & 4+1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 5 & 5 \end{pmatrix}$$

has determinant zero.

Once we have found the values of z the problem of finding the corresponding vectors c is very easy. Suppose we take the value z = 6. We are looking for a vector satisfying

$$(T-Z)c = 0.$$

If the coordinates of c are r and s, we wish to find r and s such that

$$\begin{pmatrix} -5 & 2 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} -5r + 2s \\ 5r - 2s \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

An obvious solution of this equation is r = 2, s = 5. (Any multiple of this vector will also be a solution as well. A rule of thumb for finding the solution is to take a row of the matrix of T-Z, read it from right to left and change one sign. Thus in

$$\begin{pmatrix} -5 & 2 \\ 5 & -2 \end{pmatrix}$$

we took the bottom row, changed the -2 to 2 and read from right to left to obtain (2,5). If one row consists of 0, 0 then the other row must be used. If both rows are 0 so that the matrix of T-Z is the zero matrix then any values of r and s will do.

Let us now check that if we take $z = 6$ and $c = 2a + 5b$ then

$$Tc = 6c.$$

This reduces to checking that

$$\begin{pmatrix}1 & 2\\ 5 & 4\end{pmatrix}\begin{matrix}2\\ 5\end{matrix} = 6\begin{pmatrix}2\\ 5\end{pmatrix}$$

To do this, we simply multiply out, obtaining

$$\begin{pmatrix}1 & 2\\ 5 & 4\end{pmatrix}\begin{pmatrix}2\\ 5\end{pmatrix} = \begin{pmatrix}1\times 2 & + & 2\times 5\\ 5\times 2 & + & 4\times 5\end{pmatrix} = \begin{pmatrix}12\\ 30\end{pmatrix} = \begin{pmatrix}6\times 2\\ 6\times 5\end{pmatrix} = 6\begin{pmatrix}2\\ 5\end{pmatrix}$$

Similarly, let us take the other value of z given by $z = -1$. The corresponding matrix T - Z becomes

$$\begin{pmatrix}1 + 1 & 2\\ 5 & 4 + 1\end{pmatrix} = \begin{pmatrix}2 & 2\\ 5 & 5\end{pmatrix}$$

so we take c = -2a + 2b. We must check that for this choice of the vector c we get

$$Tc = -c.$$

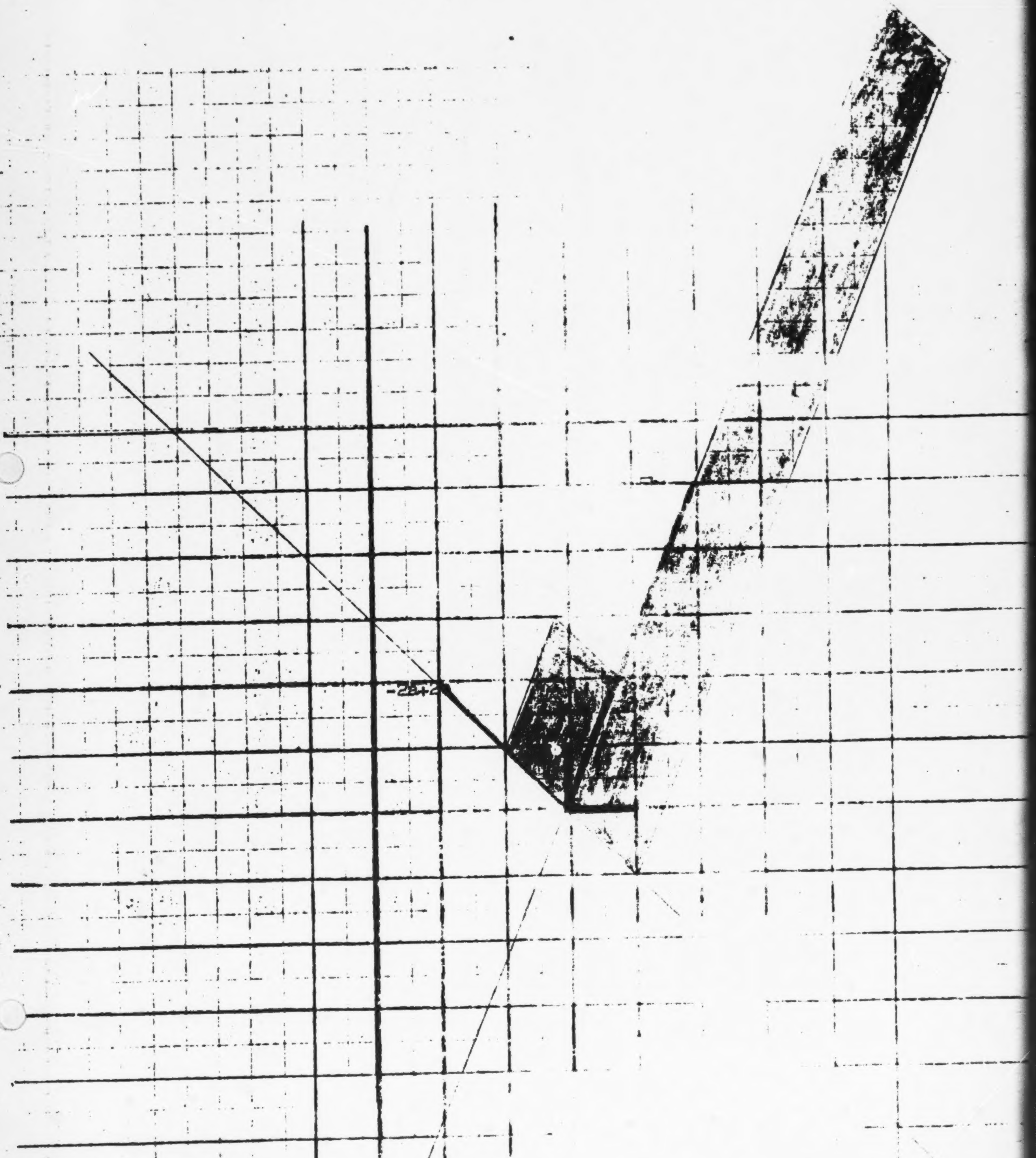
We verify this by the computation

$$\begin{pmatrix}1 & 2\\ 5 & 4\end{pmatrix}\begin{pmatrix}-2\\ 2\end{pmatrix}=\begin{pmatrix}1\times(-2) + 2\times2\\ 5\times(-2) + 4\times2\end{pmatrix}=\begin{pmatrix}-2 + 4\\ -10 + 8\end{pmatrix}=\begin{pmatrix}2\\ -2\end{pmatrix}= -\begin{pmatrix}-2\\ 2\end{pmatrix}$$

These special numbers, 6 and -1 associated to the matrix 1 2 are called the eigenvalues of the matrix. The corresponding vectors 2a + 5b and -2a + 2b are called eigen vectors of the matrix.

If we have found eigenvalues and two distinct eigenvectors (not lying on the same line) for a given matrix, then it is very easy to describe the geometric behavior of the corresponding transformation. We simply draw the lines containing these two vectors. The transformation then stretches (or contracts with possible reverse in direction) along thedirections parallel to these two lines, by the amount indicated by the eigenvalues. Thus, in the next diagram, we illustrate how the transformation T, whose matrix is given by $\begin{pmatrix}1 & 2\\ 5 & 4\end{pmatrix}$ in terms of the given basis a and b. We first draw the vectors 2a + 5b and -2a + 2b. We can then indicate how the transformation operates by drawing the image of a parallelogram to the axes provided by these eigenvectors.

The transformation carries the smaller parallelogram [illegible] one. It reflects the whole plane through the line [illegible] 2a + 5b , and stretches the plane by a factor [illegible] this line.



We now state the general method, illustrating it by another numerical example. We start with a given transformation whose matrix, in terms of our basis is

general method	specific example
$\begin{pmatrix} x & u \\ y & v \end{pmatrix}$	$\begin{pmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{pmatrix}$

We first look for the eigenvalues of the matrix. We thus wish to find those values of z for which the determinant of the matrix corresponding to T - Z becomes equal to zero;this means solving the quadratic equation

$$(x-z)(v-z) - uy = 0 \qquad\qquad (6\frac{1}{2} - z)(-4 - z) - 9\,(-3) = 0$$

or, rearranging the terms, this is the same as the quadratic equation

$$z^2 - (x + v)\,z + (xv - uy) \qquad\qquad z^2 - 2\frac{1}{2}z + 1 = 0.$$

(Notice that this quadratic equation takes on a very simple form in terms of the original matrix. The coefficient of z^2 is always one, the coefficient of z is always -(the sum of the terms on the diagonal) and the constant term is always the determinant).

We next solve this equation for z. We can use the formula for the solution of a quadratic equation,

$$z = \frac{(x+v) \pm \sqrt{(x+v)^2 - 4(xv-uy)}}{2} \qquad z = 2, \qquad z = \tfrac{1}{2}$$

Notice that there will be no real solutions if the expression under the square root sign is negative. Thus, if the expression under the square root sign is negative there will be no eigenvalues. This expression can be simplified,

$$(x+v)^2 - 4(xv-uy) =$$

$$x^2 + 2xv + v^2 - 4xv + 4uy =$$

$$(x-v)^2 + 4uy.$$

Therefore, if $(x-v)^2 + 4uy$ is a negative number, there will not be any eigenvalues of eigenvectors. Let us first restrict our attention to the case where this expression is greater than zero so that there will exist two distinct values of z given by the above formula. We shall call them z_1 and z_2 to avoid having to carry the complicated formula with us during our computations. We will return to study the case when the expression is negative or zero in a later section.

For each of the two values of z so obtained we form the matrices of the transformations T - Z which are

$$\begin{pmatrix} x-z_1 & u \\ y & v-z_1 \end{pmatrix} \text{ and } \begin{pmatrix} x-z_2 & u \\ y & v-z_2 \end{pmatrix} \qquad \begin{pmatrix} 4\frac{1}{2} & 9 \\ -3 & -6 \end{pmatrix} \qquad \begin{pmatrix} 6 & 9 \\ -3 & -4\frac{1}{2} \end{pmatrix}$$

For each of these matrices, we take a non-zero row, and read it backwards changing on sign, and this gives us the corresponding eigenvectors. Any non-zero multiple of an eigenvector is again an eigenvector. We may use this fact to obtain an eigenvector whose coordinates have a simpler looking form: The corresponding eigenvectors are thus

$-ua + (x-z_1)b$		$-ua + (x-z_2)b$	$-9a + 4\frac{1}{2}b$	$4\frac{1}{2}a - 3b$
or	and	or		
$-(v-z_1)a + yb$		$-(v-z_2)a + yb$	Multiplying this eigenvector by $-2/9$ we get the simpler looking eigenvector $2a - b$.	Multiplying by $-2/3$ we get the simpler looking eigenvector $-3a + 2b$.

The reader should check that these are indeed eigenvectors corresponding to the eigenvalues 2 and ½.

The reader should compute the eigenvalues and eigenvectors of the following matrices. Not all answers will come out looking nice.

1. $\begin{pmatrix} 0 & 2 \\ -6 & 7 \end{pmatrix}$
2. $\begin{pmatrix} -7 & 2 \\ -21 & 7 \end{pmatrix}$
3. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
4. $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$
5. $\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$
6. $\begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix}$

Change of basis. In the last section, we saw that a linear transformation having distinct eigenvalues has a very simple geometrical description in terms of its eigenvectors. It also has a very simple matrix form if we were fortunate enough to choose our basis to consist of eigenvectors. In fact, if our basis consists of eigenvectors, then the matrix is in diagonal form, that is, it takes the form

$$\begin{pmatrix} z_1 & 0 \\ 0 & z_2 \end{pmatrix}$$

where z_1 and z_2 are the eigenvalues of the linear transformation. If we were not so fortunate as to choose the eigenvectors as our basis elements, then the expression for our matrix might appear quite complicated. We should therefore study the problem of how the matrix of a linear transformation changes when we decide to make a change of our choice of basis vectors. In pursuing this question, we will be able to reformulate the results of the previous section in a form which will allow us to handle the case where the quadratic equation we encountered last time does not have real roots.

Let T be a linear transformation whose matrix, in terms of the basis a and b, is

$$\begin{pmatrix} x & y \\ y & v \end{pmatrix}$$

Suppose now that we wish to change our basis to a different choice of basis, say a' and b' where

$$a' = 2a - b, \qquad b' = -3a + 2b.$$

(Here, as usual, we have chosen specific values for a' and b' for illustrative purposes.) Let us see how to express the matrix of T in terms of the new choice of basis. We compute Ta' and Tb' by our usual procedure.

$$\begin{Bmatrix} x & u \\ y & v \end{Bmatrix} \begin{Bmatrix} 2 \\ -1 \end{Bmatrix} = \begin{Bmatrix} 2x - u \\ 2y - v \end{Bmatrix} \qquad \begin{pmatrix} x & u \\ y & v \end{pmatrix} \begin{pmatrix} -3 \\ 2 \end{pmatrix} = \begin{pmatrix} -3x + 2u \\ -3y + 2v \end{pmatrix}$$

This gives

$$Ta' = (2x - u)a + (2y \; v)b \qquad Tb' = (3x + 2u)a + (3y + 2v)b$$

However, we are interested in the expression of Ta' and Tb' in terms of a' and b', not in terms of a and b. For this we use the fact that

$$a = 2a' + b' \qquad b = 2b' + 3a'$$

which we derive by computing the inverse of the matrix

$$\left\{\begin{matrix} 2 & -3 \\ -1 & 2 \end{matrix}\right\}.$$

and verify directly. The final expression for what T does to a' and b' is as follows:

$$\begin{aligned} Ta' &= (2x - u)a + (2y - v)b \\ &= (2x - u)(2a' + b') + (2y - v)(2b' + 3a') \\ &= \left\{2\,(2x - u) + 3\,(2y - v)\right\} a' + \left\{(2x - u) + 2\,(2y - v)\right\} b'. \end{aligned}$$

Also

$$\begin{aligned} Tb' &= (-3x + 2u)a + (-3y + 2v)b \\ &= (-3x + 2u)(2a' + b') + (-3y + 2v)(2b' + 3a') \\ &= \left\{2(-3x + 2u) + 3(-3y + 2v)\right\} a' + \left\{(-3x + 2u) + 2(-3y + 2v)\right\} b' \end{aligned}$$

Thus the matrix of T, in terms of a' and b' is

$$\begin{pmatrix} 2(2x - u) + 3(2y - v) & 2(-3x + 2u) + 3(-3y + 2v) \\ (2x - u) + 292y - v) & (-3x + 2u) + 2(3y + 2v) \end{pmatrix}.$$

This complicated looking matrix can be written in a similar form which is easier to remember and understand. It can be written as the triple product

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x & u \\ y & v \end{pmatrix} \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} .$$

as can be checked by multiplying this out. The matrix

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$$

is just the inverse of the matrix

$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} .$$

Thus the rule for changing the matrix when we change the basis is the following: first write out the new basis in terms of the old basis, and write down the corresponding matrix. Let us call this the change of basis matrix. Compute the inverse of this matrix. Then the matrix of the linear transformation in terms of the new basis is given by

$$\begin{pmatrix} \text{matrix} \\ \text{in new} \\ \text{basis} \end{pmatrix} = \begin{pmatrix} \text{inverse of} \\ \text{change of} \\ \text{basis matrix} \end{pmatrix} \times \begin{pmatrix} \text{old} \\ \text{matrix} \end{pmatrix} \times \begin{pmatrix} \text{change of} \\ \text{basis} \\ \text{matrix} \end{pmatrix}$$

For example, let us take the linear transformation whose matrix in terms of a and b is

$$\begin{bmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{bmatrix}$$

According to our computations of the previous section, we know that this matrix has eigenvalues 2 and $\frac{1}{2}$ with corresponding eigenvectors 2a - b and -3a + 2b. If we set a' = 2a - b and b' = -3a + 2b as our new basis vectors, we know that the matrix of T with respect to these new basis vectors is the diagonal matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

We can verify that this coincides with our rule as formulated above. The matrix for the change of basis from a and b to a' and b' is

$$\begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}$$

while the inverse of this matrix is

$$\begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}.$$

We then check by multiplication of the matrices that

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{pmatrix} \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 4 & -1\frac{1}{2} \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Thus

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{pmatrix} \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}$$

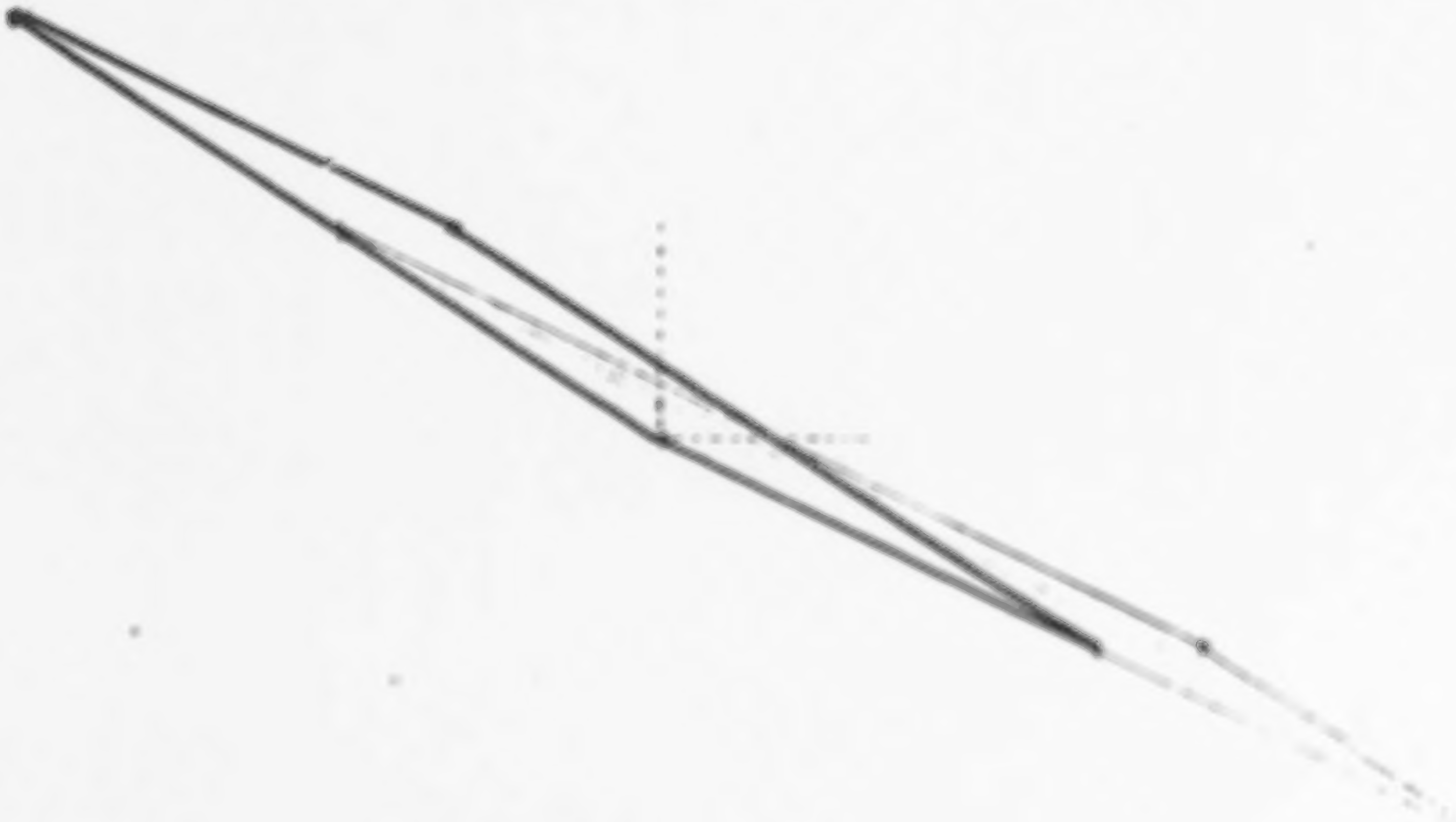
We can, of course, rewrite this last equation as

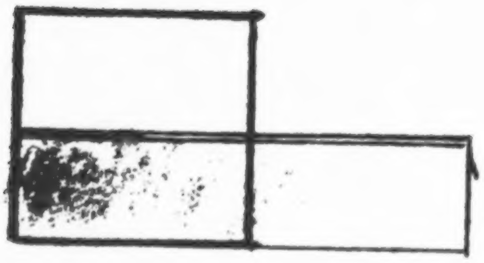
$$\begin{pmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$$

In this form, we can give a more geometrical interpretation to the equation, by considering all the matrices on the right as matrices of linear transformations. The general idea is as follows:
Suppose that we really have a preferred choice of basis vectors a and b. For instance, since we can always buy cross-section paper, it would save us a lot of work if we choose a and b to be along the axis of the cross-section paper and one inch in length. In this way, the printed lines provide us with an immediate way of computing the coordinates of any

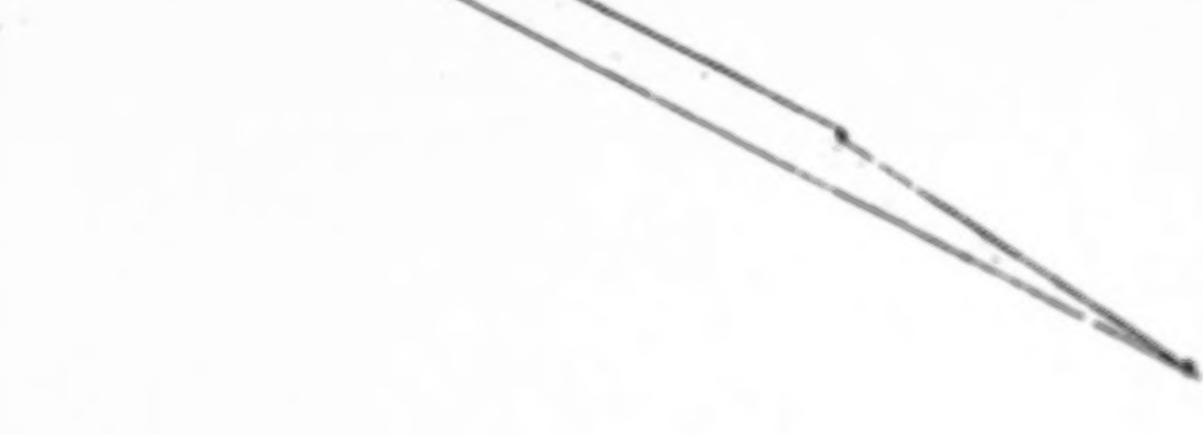
point on the plane. Now a diagonal matrix such as $\begin{matrix} 2 & 0 \\ 0 & \frac{1}{2} \end{matrix}$ has a very simple interpretation. It says, expand horizontally by a factor of two and compress vertically by a factor of one half.

Now the transformation T whose matrix is $\begin{pmatrix} 6\frac{1}{2} & 9 \\ -3 & -4 \end{pmatrix}$ expands and compresses in the directions of its eigenvectors . Thus a picture of its action is



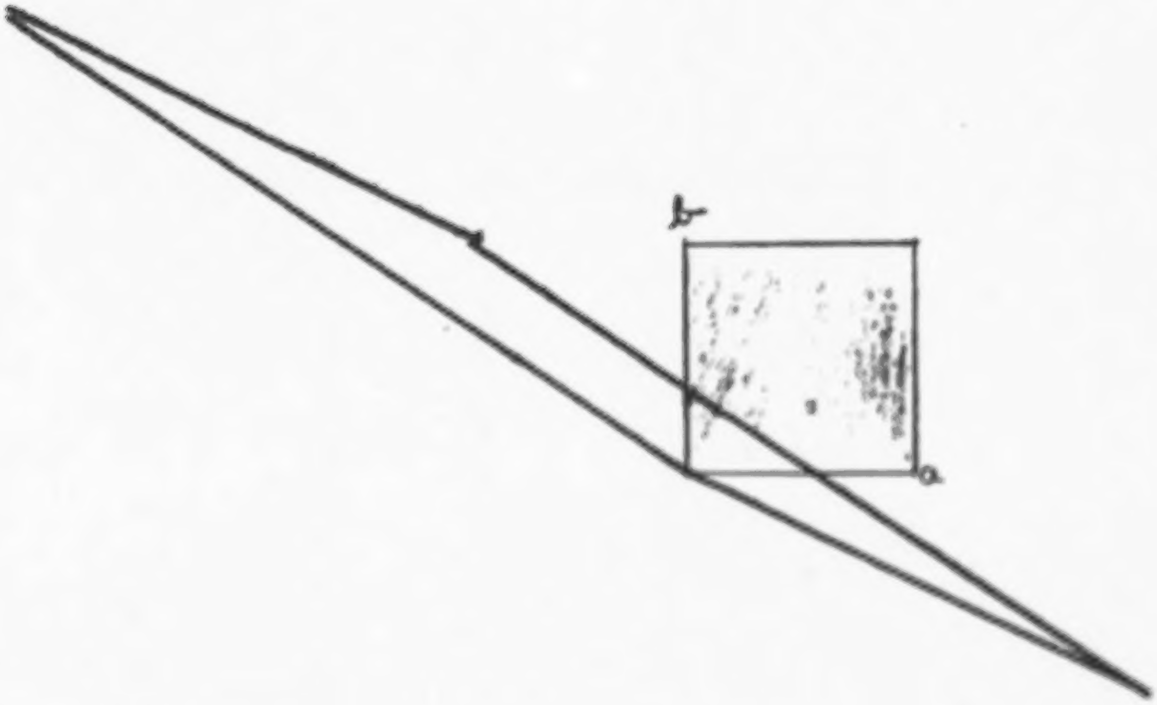


$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$



$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}$$

We can regard this action as the composition of three transformations: First "unhinge" the plane and map it into itself in such a way that the eigenvectors 2a-b and -3a + b are carried into our basis vectors a and b. Then apply the simple vertical and horizontal compression and expansion described by the diagonal matrix. Then map the plane back into itself in such a way that the vectors a and b go back into the eigenvectors again. Pictorially, we are regarding T as the composition of the three steps drawn below.



$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$$

We can thus regard a transformation of the distinct real eigenvalues as a distorted version of a diagonal transformation, whereby a diagonal transformation we mean a transformation whose matrix is diagonal in our preferred coordinate system.

The collection of all diagonal matrices has some very nice properties. Any two diagonal transformations commute,

$$\begin{Bmatrix} x & 0 \\ 0 & v \end{Bmatrix} \begin{Bmatrix} x' & 0 \\ 0 & v' \end{Bmatrix} = \begin{Bmatrix} x' & 0 \\ 0 & vv' \end{Bmatrix} = \begin{Bmatrix} x' & 0 \\ 0 & v' \end{Bmatrix} \begin{Bmatrix} x & 0 \\ 0 & v \end{Bmatrix}$$

A diagonal matrix has an inverse if and only if both entries along the diagonal are not zero, in which case the multiplicative inverse is given by the formula

$$\begin{Bmatrix} x & 0 \\ 0 & v \end{Bmatrix}^{-1} = \begin{Bmatrix} 1/x & 0 \\ 0 & 1/v \end{Bmatrix}$$

Conformal transformations. We now wish to study linear transformations which do not possess real eigenvalues; thus those linear transformations whose matrix

$$\begin{Bmatrix} x & u \\ y & v \end{Bmatrix}$$

satisfies

$$(x-v)^2 + 4uy < 0$$

We shall describe a "nice" collection of linear transformations with this property. We shall see later on that any linear transformation with $(x-v)^2 + 4uy \quad 0$ can be regarded as a distortion of one of these transformations by a skew choice of basis.

Consider a transformation whose matrix has the form

$$\begin{pmatrix} r & s \\ -s & r \end{pmatrix}$$

In this case $x = r$ and $v = r$ while $u = s$ and $y = -s$ so that the expression

$$(x-v)^2 + 4uy = -4s^2 \leq 0$$

In experiments the following facts are brought out concerning transformations of this type.

Any transformation of this type is a similarity transformation of the plane, that is, it carries any figure into a figure similar to it.

It distorts length by a factor of $\sqrt{r^2 + s^2}$, and can be regarded as the composition of a rotation of the plane followed by the transformation that changes all distances by the amount $\sqrt{r^2 + s^2}$.

A similarity transformation is sometimes called a linear conformal transformation. A conformal transformation preserves angles but need not preserve lengths. As a convenient convention, we shall agree to call the zero transformation conformal.

Let S and T be two conformal transformations. If either S or T is zero, then clearly S T is zero. If both S and T are not zero, then T and S both preserve angles. If we first apply T and then apply S we will still have preserved all angles. Thus S T will again preserve angles. Thus the composite of two conformal transformations is again conformal.

Let us check this fact by looking at the product of the corresponding matrices. Suppose that T has the matrix $\begin{matrix} a & b \\ -b & a \end{matrix}$ and S has the matrix $\begin{matrix} u & v \\ -v & u \end{matrix}$. Then multiplying the matrices gives

$$\begin{pmatrix} u & v \\ -v & u \end{pmatrix}\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \begin{pmatrix} au-bv & bu+av \\ -bu-av & au-bv \end{pmatrix}.$$

We see that the product matrix has the form

$$\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$$

where

$$x = au - bv \quad \text{and} \quad y = bu + av$$

and thus corresponds to a conformal transformation.

Notice that the sum of two conformal matrices is again conformal. Indeed

$$\begin{pmatrix} u & v \\ -v & u \end{pmatrix}+\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \begin{pmatrix} (u+a) & (b+v) \\ -(b+v) & (u+a) \end{pmatrix}$$

and the matrix on the right has the desired form.

Let us go back to the expression for the product of two conformal matrices:

$$\begin{pmatrix} u & v \\ -v & u \end{pmatrix}\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \begin{pmatrix} au - bv & ub + va \\ -(bv + bu) & au + bv \end{pmatrix}$$

Let us now multiply tham in reverse order:

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} u & v \\ -v & u \end{pmatrix} = \begin{pmatrix} au - bv & av + bu \\ -(bu + av) & -bv + au \end{pmatrix}$$

Notice that we get the same answer. Thus, if S and T are conformal linear transformations, we have

$$S \circ T = T \bullet S$$

i.e., multiplication is commutative.

When will a conformal transformation have a multiplicative inverse? If T is a conformal linear transformation whose matrix is

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

we know how to answer this question. We must check the determinant of this matrix which is

$$a \times a - b \times (-b) = a^2 + b^2 .$$

Notice that this expression can be zero only when a and b are both zero. Thus if T is a non-zero conformal linear transformation, it possesses a multiplicative inverse.

Notice that the collection of all conformal matrices behaves a lot like the number system. We have noticed on page that some of the laws for numbers break down for the collection of all matrices. The commutative law does not hold for the collection of all matrices. It *does* hold for the collection of all coformal matrices. For the collection of all matrices, it is *not true* that any nonzero element has a multiplicative inverse. For the collection of conformal linar transformations it *is* true that every non-zero element has a multiplicativ inverse.

Let us now examine two special conformal transformations. The first one is our old friend the identity transformation. The identity transformation clearly preserves angles, because it preserves all geometric figures; it does not move the plane at all. Remember that its matrix is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The second transformation we wish to consider is the one whose matrix is

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

It is clearly conformal. What does it correspond to geometrically? It changes the length of any vector by a factor $0^2 + 1^2$ - that is it doesn't change length at all. It sends the vector (1,0) into the vector (0, -1), that is, it rotates it clockwise by ninety degrees. Similarly, it sends the vector (0, 1) into the vector (1,0) - that is, it rotates it also by ninety degrees. Therefor we conclude that it rotates every vector in the plane by ninety degrees as can be checked in the experiments.

Rotating through ninety degrees twice is the same as rotating through one hundred and eighty degrees. We can check this fact directly by matrix multiplication:

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{matrix} -1 & 0 \\ 0 & -1 \end{matrix} \quad .$$

We can write this last equation as

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Now the point of this discussion is that we can express any conformal matrix in terms of the two special matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

In fact,

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Let us give special names for these two conformal transformations: let us call

$\mathbf{1}$ the identity form

and i the transformation whose matrix is

We can thus write every conformal transformation of the place as

$$a\mathbf{1} + bi \quad .$$

We know that multiplication of two conformal linear transformations is commutative and the rest of the usual laws of multiplication hold. The transformation is the identity for multiplication while

$$i \bullet i = i^2 = -\mathbf{1}$$

We can use these rules to reconstruct the rule for multiplication of conformal linear transformations. Suppose we wish to multiply

$$(a\mathbf{1} + bi) \times (u\mathbf{1} + vi)$$

We get,

$$(a\mathbf{1} + bi) \times (u\mathbf{1} + vi) = (a\mathbf{1}) \times (u\mathbf{1}) + (a\mathbf{1}) \times (vi) + (bi) \times (u\mathbf{1}) + (bi) \times (vi)$$

by the distributive laws. Since $\mathbf{1}$ is the identity for multiplication we have

$(a\mathbf{1}) \times (b\mathbf{1}) = a\mathbf{1}$, $(a\mathbf{1}) \times vi = av\,i$, $bi \times a\mathbf{1} = bu.i$

and since $i^2 = -1$ we have

$$(bi) \times (vi) = -bv\mathbf{1}$$

thus

$(a\mathbf{1} + bi) \times (u\mathbf{1} + vi) = (a \quad -bv)\mathbf{1} + (av + bu)i$ whose matrix is

$$\begin{pmatrix} au - bv & av + bu \\ -(av + bu) & au - bv \end{pmatrix}$$

If we compare this with the equation on page[illegible] we see that we have obtained the same answer. Since acts as the identity for multiplication, the product

$(a\mathbf{1}) \times T$ is the same as aT

for any transformation T. Thus as far as multiplication is concerned, we can suppress the $\mathbf{1}$ in making computations. For this reason, people usually write

$$a + bi \quad \text{instead of} \quad a\mathbf{1} + bi \; .$$

When we write conformal linear transformations this way, they are called complex numbers. Thus a complex number is an expression of the form

$$a + ibi$$

where all the rules of arithmetic apply, together with the rule $i^2 = -1$.

5

LABORATORY MANUAL FOR CHAPTER 3

VECTORS IN THE PLANE

The purpose of the next collection of experiments is to study properties of translations in the plane. The equipment consists of cross-sectioned graph paper ruled 8 squares to the inch and plastic transparency sheets also ruled 8 squares to the inch. In addition you will also use your straight edge and compass. The purpose of the ruling on the paper and on the plastic transparency sheets is to insure that no rotation occurs while sliding it along the paper.

The first two experiments demonstrate that a directed segment determines a translation and two directed segments determine the same translation if and only if they are parallel, of equal length, and point in the same direction.

The next three experiments show the geometric meaning of the group laws (commutative and associative laws).

We then study scalar multiplication and vector space properties the notion of basis, lattice and so on.

EXPERIMENT I TRANSLATIONS

Directions

1. Draw a directed segment a_1, a_2, anywhere on the ruled paper.

2. Line up the plastic transparency over the ruled paper making sure the horizontals line up with the horizontals and the verticals line up with the verticals.

3. Mark the point, a_1, with a felt tipped pen, lying over the initial point of the directed segment.

4. Now slide the transparency to the position with the marked point lying over the terminal point of the directed segment. This is the translation associated with the directed segment.

5. Return the transparency to its original position; that is return the marked point to the initial point, a_1. Puncture the transparency at some other point, marking the ruled paper underneath. Call this point b_1. Now slide the transparency so that the point originally lying over a a_1 now lies over a_2. Mark the position of the hole on the ruled paper. Call this point b_2.

6. Check that the line determined by b_1 and b_2 is parallel to the line determined by a_1 and a_2. Check that the segment b_1, b_2, has the same length as a_1, a_2 and points in the same direction.

7. Repeat 4, 5, and 6 with some other puncture.

EXPERIMENT II EQUIVALENT DIRECTED SEGMENTS

Steps 4, 5, and 6 of the previous experiment show that a translation together with a "starting point" b_1 determine an "ending point" b_2. We say that the translation transforms b_1 into b_2. We say that the directed segment b_1 b_2 corresponds to the translation. We saw that if two segments a_1 a_2 and b_1 b_2 correspond to the same translation then they lie on parallel lines, are of equal length and point in the same direction.

We now establish the converse.

Directions

1. Draw a directed segment a_1 a_2.
2. Choose any point b_1 on the ruled paper.
3. Draw the line through b_1 which is parallel to the line determined by a_1 and a_2.
4. Mark the point b_1 on this line so that b_1 b_2 has the same length as a_1 a_2 and points in the same direction.
5. Place the transparency over the ruled paper and mark, with felt tipped pan, the points lying over a_1 and b_1.
6. Slide the transparency so that the point originally over a_1 now is over a_2. Check that the point originally over b_2 now lies over b_2.

We say that a_1 a_2 and b_1 b_2 are equivalent directed segments.

Two equivalent directed segments determine the same translation. The translation is called a vector. We use the letter v to denote a vector. We sometimes also use the letter T_v to emphasize that we are thinking of v as a motion of the plane. (The letter T stands for transformation).

EXPERIMENT III ADDITION

Let T_{v_1} and T_{v_2} be two translations. Let $T_{v_2} \circ T_{v_1}$ be the transformation obtained by first applying T_{v_1} and then T_{v_2}. We check that $T_{v_2} \circ T_{v_1}$ is again a translation. We denote it by T_{v_3} and we write $v_3 = v_1 + v_2$. We check that $v_1 + v_2$ $v_2 + v_2$.

Directions

1. Draw two directed segments $a_1\ a_2$ and $b_1\ b_2$ on the ruled paper.
2. Place the transparency over the ruled paper, line it up, and puncture the points over $a_1\ a_2\ b_1$ and b_2.
3. Slide the transparency until the point corresponding to a_1 is over b_2, line up the transparency and mark the point which was over a_2 on the ruled paper. Call this point **c**.
4. Place the point corresponding to b_2 over a_2 mark the point corresponding to b_2 on the ruled paper. Call this point d.
5. Notice that $b_2 0$ and $a_1 d$ are equivalent, they determine the same translation. This experiment is more striking if we choose our segments with the same intial point.
6. Let 0a and 0b be two directed segments under the same initial point 0. Find the sum of the corresponding vectors and demonstrate the commutative law.

Recall that the translation corresponding to a directed segment a_1 a_2 is called a vector. Frequently, since any vector is merely a representative of equivalence class of segments, we choose a common initial point 0 to represent all vectors. This point is called the origin. Since now all segments have the same initial point we can identify them by merely noting their terminal point. We will denote the vector corresponding to 0a by a and will therefore denote the endpoint of the segment corresponding to a + b and whose initial point is 0 by a + b.

EXPERIMENT IV THE ASSOCIATIVE LAW

Directions

1. Choose 0 and three points a, b, and c.
2. Construct a + b and b + c.
3. Construct (a + b) + c and a + (b + c).

Scalar Multiplication

As usual, we denote a + b by 2a and, more generally a + a+ a+ ... + a (n times) by na. We can also construct a vector $\frac{1}{2}$a as the vector satisfying $\frac{1}{2}$a + $\frac{1}{2}$a = a. In fact we can find it by bisecting the segement Oa.

EXPERIMENT V THE DISTRIBUTIVE LAW

Directions

1. Choose vectors a and b.
2. Construct a + b.
3. Construct (1 + $\frac{1}{2}$)a by extending vector a half again as much in the same direction. Do the same for (1 + $\frac{1}{2}$)b. Construct (1 + $\frac{1}{2}$) (a+b) by extending (- b) half again as much in the same direction.
4. Add (1 + $\frac{1}{2}$)a to (1 + $\frac{1}{2}$) b.
5. Form (1 - $\frac{1}{2}$) (a + b), by extending (a + b) half again in the same direction. Notice that

$$(1 + \tfrac{1}{2})\ (a + b) = (1 + \tfrac{1}{2})a + (1 + \tfrac{1}{2})b$$

Experiment VI

Here are two vectors a and b. Construct $2a + b$.

EXPERIMENT VII INVERSE OF A VECTOR

Let a_1 a_2 be any directed segment corresponding to a vector a so that translating by a moves a_1 to a_2. Then the inverse of a will move to a_2 to a_1. That is a_2 a_1 represents -a.

Construction of -a from a fixed origin.

Directions.

1. Choose point a.
2. Puncture the transparency over 0 and a, after lining up the grids.
3. Place the point which was over a over the origin 0 and line up the grids.
4. Mark the point which was formerly over 0 on the ruled paper. This is -a.

EXPERIMENT VIII

Directions

1. Choose vectors a and b.
2. Construct the following points:

   2a, a+b, 2b, 2a-b,2a-2b, a-b, a-2b,-a, -2a, -b, -2b, -a-b.

   -2a-b, -a-2b, -2a-2b, b-a, b-2a, and 2b-2a.

EXPERIMENT IX - THIS EXPERIMENT REQUIRES A LARGE SHEET OF PAPER WHICH WILL BE USED REPEATEDLY IN THE SUCCEEDING EXPERIMENTS.

1) Choose an origin O near the center of the paper.
2) Choose points a and b so that Oa and Ob are not too large or too small, say somewhere between one and three inches each.
3) Form the vectors a, 2a, 3a, 4a, 5a, -a -2a, -3a, -4a, -5a and the vectors b, 2b, 3b, 4b, 5b, -b, -2b, -3b, -4b, -5b. The easiest way to draw these vectors is by drawing the line through O and a and then marking off the points with a compass. Similarly for b.
4) Form the point -5a + b. Draw the line through -5a and -5a+b and then mark off the points -5a+b, -5a+2b, -5a+3b, -5a+4b -5a+5b and -5a-b, -5a-2b, -5a-3b -5a-4b and -5a-5b.
5) Construct 5b+a. Draw the line through 5b and 5b+a and then mark off the points 5b+a, 5b+2a, 5b+ 3a, 5b+4a, 5b+5a, 5b-a, 5b-2a, 5b-3a, 5b-4a and 5b-5a
6) Draw the line passing through -5a_5b and -5a, the line through -4a+5b and -4a until the line through 5a+5b and 5a. There are eleven lines in all

7) Similarly, draw the eleven lines through -5a+5b and 5b, -5a+4b and 4b etc., eleven lines in all.
8) Label many of the points of intersection such as a+b, a+2b, 3a-4b etc. In order to same time and space use the following notation; (3,4) stands for 3a+4b, (1,1) stands for a+b, (-2,3) stands for -2a+3b and so on.
9) Get some practice with this notation. Locate the points (3,-3), (2,-2), (1,1 the points (4,2), (2,1), (-2,-1), the points (4,0), (2,0), (-3,0).

We have subdivided the plane into parrallelograms. We can use one corner of the parrallelogram to label the parrallelogram. We use the point

(3,-4) to label the parallelograms whose corners are 3a-4b, 4a-4b, 3a-3b and 4a-3b .

10) Which is the parallelogram labeled by (2,-3) on your sheet?

EXPERIMENT X

Use the same sheet as in experiment IX

1) Find the point $3\frac{1}{2}a + 2\frac{1}{2}b$.

2) Find the point $-\frac{1}{2}a + b$.

EXPERIMENT XI

1) Pick a point Q. In which parallelogram does it lie? What is the label of this parallelogram.

2) Subdivide the parallelogram containing Q into four congruent parallelograms. What is the label of the small parallelogram containing Q?

3) Subdivide the small parallelogram containing Q once more into four. What is the label of the parallelogram containing Q?

EXPERIMENT XII

Continue to use the same sheet as in experiment IX . Draw all the figures of this experiment in very light pencil as they will not be used in the next experiments, while the sheet itself will continue to be used.

1) Through the point Q of experiment XI draw the line parallel to the line through 0 and b .

2) Mark the point of intersection of this line with the line through 0 and a . Call this point c.

3) The point c lies on the line through 0 and a . If we take 0 as the origin on this line and we use a as a basis of this line then we know that c = ra where r is some real number . We can find the expansion of this real number according to the methods of Chapter II . Find the expansion of r up to the second dyadic place .Compare this with the first term of the answer to part 3) of experiment XI .

4) Repeat parts 2) ,3) and 4) interchanging a and b . That is, draw the line through Q which is parallel to the line through 0 and a . Mark the point of intersection of this line with the line through 0 and b ; call this point d . Expand d in terms of b , finding the first two terms in the expansion of s where d = s b . Compare your answer with the second term of the answer to part 3) of experiment XI .

5) Observe that

$$Q = c + d$$

so that we may write

$$Q = ra + sb \quad .$$